

Environment and the public's purses

Japan's environment agency altruistically advocates sacrifice on behalf of the global environment, but should instead be seeking the basis of common self-interest.

SMOKESTACK Japan is an international image, most other people's vision of what is now the richest country in the world. Japan is generally supposed to be a place in which dirty factories compete with people for room in which to site themselves, a vivid demonstration that technology is inimical of quality of life. The reality is, of course, quite different. In less than fifteen years, Japanese cities have been transformed by stringent regulation of air and water quality. The national talent for flower-arranging has been spectacularly transplanted to city sidewalks. The richest country in the world belies the stereotyped image, although in part because so much of it is spanking new.

Yet Japan's environmental agency is now (see page 385) demanding further environmental measures, some of them likely to be beneficial for those on whom the costs will fall, others more altruistically aimed at the general improvement. Even if the agency may carry little influence with the powerful economic ministries in Tokyo, so that many of its proposals will be delayed or overlooked, environmentalists elsewhere have much to learn from the circumstance that this altruism is so closely linked with Japan's comparatively new riches.

The first stark lesson to be drawn is that environmental quality must be purchased, and that a wealthy people can purchase more of it than others. There is nothing novel in that. There is a wealth of experience to show that communities more easily rid themselves of bacterial infection (one of the primitive environmental hazards) as they become more prosperous. The provision of an artificial environment, generally known as a house and widely believed to be a protection against the dangers of the natural environment, is positively correlated with prosperity. And can it be accidental that California, for many decades among the most well-to-do of the United States, has so often been in the van of environmental movements, many of them sensible? Or that the poor countries of the world, both the heavily-industrialized countries of eastern Europe and the agrarian countries of the developing world, so often run into environmental scrapes? That implies that industrial development, the most economical way of generating wealth, is a prerequisite of environmental quality. So why is it that the relationship between industry and environment is so often supposed to be inimical?

Part of the explanation is the general confusion about the notion of environmental quality, which means altogether too many different things. Clean air and clean water are generally understood to be public goods which are ultimately purchased in industrialized countries by consumers, either through taxes or by paying the higher prices for the goods they buy which are required to compensate manufacturers for meeting more stringent standards. There may be arguments about the equitability with which these costs are distributed among people, or about the point at which the balance should be struck between the degree of environmental quality and its cost, but a coherent society can in principle reach rational decisions. In these simple regards, most reasonably prosperous societies are probably at present purchasing what environmental quality they can afford — and hoping to buy more as time goes on.

Several of the items on the Japanese agency's shopping list

present different problems because they are international. On the assumption, safe enough, that chlorofluorohydrocarbons are inimical of stratospheric ozone, the most easily identifiable losers from their accumulation are people living at high altitudes and low latitudes, not the Japanese. That is the sense in which it was altruistic of Japan (and other similarly placed governments) to have subscribed to last year's convention on chlorofluorohydrocarbons — they have made a public purchase whose benefits will disproportionately accrue to others. There are many, of course, who say that the Montreal protocol is insufficient — that the public purchase has been too little — but that view overlooks the unavoidable limitations of altruism in the behaviour of governments. It is as valid to regard the Montreal protocol as a resounding triumph of the general interest over conflicting national self-interests — and to look for mechanisms by which the conflicts may be blunted in the future.

The carbon dioxide problem, although its consequences are potentially much more serious than those of diminished stratospheric ozone, may paradoxically be more simply tackled. For the greenhouse effect, if its signal emerges from the climatic noise in the years ahead, should create an identity of interest among very different governments. Poor countries will be particularly fearful of drastic climatic change, especially because predictions are so uncertain. Rich countries (among which only Switzerland is landlocked) would all suffer from the increase of sea level that could follow increased surface temperature over many years. Moreover, the cost of purchasing relief from the greenhouse effect (dispensing with most machines dependent on fossil fuel) would be so great as to require investment over decades. It is not in the circumstances unreasonable to ask that governments should tackle a task of a character they inherently detest — that of answering the hypothetical question "What will we jointly do if...?"

From past and present experience, especially in Japan, a large part of the answer must be "Create more wealth". If, for example, it should be necessary to replace existing electricity generating plants with safely operated nuclear plants, there will be a need of capital investment on a breathtaking scale. But it will also be necessary for governments to be more discriminating than at present about the purchases they make of environmental quality. The present row in Britain about the preservation of the physical appearance of southeast England, a rolling man-made landscape spotted with low-density housing, echoed as it is in many other prosperous communities, could easily turn out to seem a frivolous luxury. Even Japan's environment agency's breast-beating about the logging roads Japan helps to build in Malaysia may seem an irrelevance if that is the cheapest way of helping Malaysia to become more prosperous.

What this implies is that the supposed conflict between economic growth and environmental quality is not merely an illusion but a dangerous trap, into which too many people have repeatedly tumbled. Japan's environment agency, for all its altruism, seems insufficiently aware of the danger — and of the economic foundations of its own environmental achievements in recent years. Would it have more influence with the economic ministries if it learned this lesson? □



Have bombs spread?

Crypto-nuclear powers may be as dangerous as the genuine article.

LAST week's dispute between Hungary and Romania about the disposal of 12.5 tonnes of Norwegian heavy water (page 384) is more than merely another sign of the growing readiness of Eastern European countries to be seen squabbling in public; it is also a reminder that the Nuclear Non-Proliferation Treaty (NPT) needs continuing attention, certainly more attention than it has been getting in the past year's excitement about other negotiations in arms control, of which the INF treaty is the tangible product so far. Much the same conclusion can be derived from the conviction last month of Mr Moredechai Vununu, an Israeli technician, of treason after a trial held in camera before a military court. The case against Vununu stemmed from his disclosure to a London newspaper (the *Sunday Times*) last year of information about Israeli operations at the Dimona reactor site in the Negev. Vununu's allegations, as eventually published, were that Dimona has a secret plant in which plutonium can be (and is) separated from reactor fuel. None of this circumstantial evidence proves that Israel is already a nuclear power, but it does suggest that familiarity — and perhaps even the welcome relaxation of tension between superpowers — is breeding sloppiness in people's regard for the danger that nuclear weapons will spread.

The case of Mr Vununu is like a problem in logic-chopping for undergraduates. There are two premises: either he was lying when he talked to the newspapers, or he was telling the truth. To lie so as to suggest that a government is making nuclear weapons when it is not may plainly be categorized as "slandering the state", but that misdemeanour does not feature in Israel's criminal code, and hardly ranks as treason, so that the conviction must be unjust. If, on the other hand, Mr Vununu was telling the truth, both the conviction and the secrecy of the trial are explicable, but then it is difficult to understand why Israel chose to hold a trial whose outcome could only reinforce the suspicion that Israel is, secretly, already a nuclear power.

There are, of course, other possibilities. Mr Vununu may have broken obligations of confidentiality towards the Israeli public service in much the way that Mr Peter Wright, the retired British intelligence officer, has offended the British government, but then one would have expected an explanation to that effect to have been forthcoming. Alternatively (requiring cynicism of an unusual degree), Mr Vununu may be neither a liar nor an honest man but a participant in a charade whose objective, after the leaking of intriguing but misleading information, a secret trial and a conviction, was to remind states in the Middle East of Israel's status as a crypto-nuclear power.

Whatever the truth, Israel's nuclear status remains much what it has been for a quarter of a century. The original Dimona reactor, a water-cooled reactor of the kind commonly called a "materials-testing reactor" in the late 1950s, and built with French assistance, could have produced enough plutonium to make one bomb a year even if its power rating had not been increased, as Mr Vununu suggested. Israel has never carried out a nuclear test, it is true, but neither has it compellingly denied that it has nuclear weapons, signed the NPT or opened Dimona to international inspection. And Israel is not alone. During the past few years, and especially since the onset of trouble in Afghanistan a decade ago, Pakistan has been talking and behaving in much the same way. So much of the technology of nuclear energy is now so familiar that a government wishing to suggest to its neighbours that it has nuclear weapons up its sleeve can do so without incurring the full expense of actually making bombs, but in the process necessarily creates the impression that nuclear weapons have indeed spread.

The governments whose interest is to prevent the spread of nuclear weapons appear not to appreciate the dangers of the

growing conviction that proliferation is not merely unavoidable but already under way. If it were otherwise, they would be more vigilant in enforcing the rules of the agreement between nuclear suppliers that bits and pieces of nuclear plants will not be sold to anybody who wishes to buy them. (It is not a joke that 15 tons of Norwegian heavy water exported in 1983 cannot now be accounted for, or that there is confusion about the earlier shipment to Romania.) More to the point, the two major powers should be more diligent in exerting the influence of which they are fond of boasting in the sense of persuading their dependent allies to toe the line of non-proliferation. There are also issues that should be dealt with within the framework of the NPT, to which almost only the crypto-nuclear powers do not belong. Why not plan to persuade France and China, both declared nuclear powers, to join the treaty by the time of the next review conference in 1990, and to find some kind of accommodation with the crypto-nuclear powers (India is the most willing to talk) by that date? It might be necessary to pay a price, say a comprehensive test-ban, but that would be no bad thing. □

Burying von Humboldt?

Extra money is only a partial remedy for West Germany's universities.

IN a misanthropic world, it is inevitable that people should take comfort in others' misfortunes. Thus academics almost everywhere will read with guilty pleasure the account on page 385 of the problems of West Germany's universities and of the remedies proposed for them by the Wissenschaftsrat. For even if many of the misfortunes confronting West German academics are in some sense general, some of them are peculiar to West Germany and also self-inflicted. But there can be no joy in a situation where it seems that, yet again, the weight of events conspires against the the marriage of research with teaching in higher education — Prussian von Humboldt's eighteenth-century legacy to the rest of Europe.

Not that the causes of the drift of researchers towards research institutes are that mysterious. The mounting cost of research equipment argues for the provision of central facilities. In West Germany and many other European systems, great national enterprises (nuclear energy, aeronautics, now scientific medicine) have created large groups of researchers with no choice but to adapt to changing circumstances by diversifying their activities, competing with universities not just for funds but people. West Germany's distinctive network of Max-Planck institutes may not be as serious a threat to the strength of university research as they may seem, given that they often crystallize around university groups or are otherwise sited intelligently. Yet the institutes assist researchers who believe teaching to be temporarily an encumbrance to escape from it for good. As competitive pressures further strengthen, the temptation to escape can only become stronger.

The Wissenschaftsrat's remedies are sensible in themselves, but are unlikely to be sufficient. More money for university research (presumably through the Deutsche Forschungsgemeinschaft or DFG) would no doubt help a little, but the federal government is already panicky about the impending decrease of taxation rates, and is unlikely to share that view. The much more serious difficulty is that the reputation of the German university system, false though it is, suggests a general air of chaos. Too many universities are too big. Too many students stay too long. Legal battles between universities and their students and/or teachers may have declined in number during the past decade, but folk-memories persist. But what most West German universities have on their side is the knowledge that they are not simply part of a national system, but also the creatures of their regional governments. After the excesses of the great expansion, some *Land* government may yet see the virtue of running a tight ship in higher education. □

First US patent on TPA for Oxford University

- Patent highlights sugar sidechains
- Other TPA patents expected

Washington

THE University of Oxford plans to announce this week that it has been granted a US patent covering the clot-dissolving glycoprotein known as tissue plasminogen activator (TPA). This development will surprise, but also disappoint, many other groups working towards this prize.

The patent is the first to be awarded for TPA in the United States, and is unique among other patents for proteins in that it specifies exactly the sugar groups attached to the protein's basic amino acid sequence.

It has been expected for some time that the first US patent for TPA would be issued this summer, but most observers had predicted that it would go to Genentech, the Californian biotechnology company that received approval to market TPA as a drug to treat heart attack at the end of last year. Genentech sells a recombinant version of TPA based on the overexpression of the human TPA gene inserted into cell cultures of the Chinese hamster ovary. The company has filed a US patent for TPA which is believed to be nearly as broad as the British patent it held until a year ago, but which was then declared invalid by a British court in a suit brought by the drug company Wellcome plc (see *Nature* 328, 189; 1987).

Genentech's British patent, now under appeal, covered all forms of human TPA produced by recombinant DNA technology, and all means for producing it.

The University of Oxford patent is not based on recombinant DNA technology. Its broadest claim is for human tissue plasminogen activator derived from cultured normal human colon cells, specifically fibroblasts. Further claims describe in detail the oligosaccharide sidechains of two forms of TPA produced by the cultures: one that acts quickly in the bloodstream to break down a blood-clot, and another that acts more slowly.

The Oxford patent is based on the premise that proteins produced by eukaryotic cells may exist in several different forms, differing in physiological activity, according to the type, number and position of the oligosaccharide sidechains added after the protein is assembled from its amino acids. There may be several different "glycoforms" of a protein within the same tissue.

Previous patent claims for either recombinant or naturally-derived proteins with potential uses as medicines have taken no account of the various glycosylation

patterns of the proteins that there may be.

The multinational pharmaceutical and chemical company Monsanto supported the research leading to the Oxford patent. Monsanto, interested in the possibility that the sugar sidechains of proteins may affect their activity as drugs, began supporting the Oxford Glycobiology Unit led by Raymond Dwek in 1983. Under the agreement, Monsanto contributes \$1.5 million a year to Dwek's unit, in exchange for the right of first-refusal to license any patentable technology. Monsanto will license the Oxford patent once it is issued, probably in the third week of this month.

Monsanto's subsidiary G.D. Searle has plans to market a form of TPA under development by Monsanto and the St. Louis biotechnology company Invitron, in which Monsanto holds a controlling interest. Invitron and Monsanto are being sued by Genentech for theft of trade secrets by two Invitron employees who formerly worked in scaling up the production of Genentech's recombinant protein products.

Genentech has charged that Invitron hired away its employees to gain access to information about its proprietary protein expression technology, with the intention of developing competing products. The suit is just beginning, but Genentech is asking for damages in the tens of millions of dollars, and a permanent injunction against sales of any forthcoming Invitron products.

Genentech and the Monsanto-Invitron collaboration are only two of many companies with various forms of TPA under development. Furthest along with modified, or second-generation, versions of TPA are Genetics Institute and Integrated Genetics. Genetics Institute's approach involves streamlining the TPA molecule by altering some of its domains to give it a longer half-life in the bloodstream, which should prevent clots from reforming after a heart attack has been stopped. The technology for altering the configuration of the sugar sidechains of a recombinant protein after it has been expressed is also being developed by Genzyme, which has filed patents on its technique.

It is believed that many of the roughly 19 companies working on TPA have filed patents covering their methods and formulations. In the United States, filed patents (known as patent applications elsewhere) are confidential until they are issued (granted), so it is difficult to say

Refusniks and *glasnost*

JEWISH "refusnik" scientists in Moscow are planning to test the limits of *glasnost* this autumn by holding an international symposium to which foreign participants will be invited. The organizers, Yuri Chernenyak, a physicist, and Igor Uspeski, an entomologist, say that they are in touch with some 60 scientists who are denied permission to emigrate, only 10 per cent of whom have jobs "more or less adequate for their qualifications". This, they say, shows there is no substance to the latest stance of Soviet officials that to permit free emigration would be to connive at a brain-drain.

Although Soviet scientists caught in the limbo between losing their jobs, after applying to leave, and receiving an exit visa have been attempting to organize international meetings since 1972, the authorities have repeatedly intervened since 1974 to prevent them happening.

Those planning this year's meeting hope to work within the terms of the May 1986 decree permitting the establishment of "informal associations and hobby-clubs". But several attempts by Jewish seminar groups to register as "informal associations" have been rejected on technical grounds. V.R.

More research centres

BRITAIN'S Science and Engineering Research Council has invited bids from higher education institutions to host a further six interdisciplinary research centres. The centres will be in process simulation, integration and control; high performance materials; optical and laser related science and technology; surface engineering; polymer science and technology; and application of parallel and novel architecture computing in science and technology. It is hoped that the new centres will start in the autumn of 1989. Meanwhile, the first research centre, in high-temperature superconductivity and based at the University of Cambridge, has appointed its director. He is Dr Peter Duncumb, who before joining the university's Earth sciences department in January was director of an industrial physics laboratory. The centre is to be formally inaugurated by education secretary Kenneth Baker on 21 June. S.H.

how many patents for TPA will emerge, and how broad their claims may be.

Issuing a patent to the University of Oxford will not prevent the US Patent Office from issuing further, perhaps even broader, patents covering TPA. But this development may set a precedent forcing other companies fully to characterize and define the glycosylation patterns of protein products, of which TPA is only one, for which they seek protection.

Carol Ezzell

Merger mooted in UK research

London

THE division of responsibilities for research and training in the biological sciences in Britain at present shared by four research councils, is to be reviewed following assertions from three of the councils that the system is unsatisfactory. Essentially, the Natural Environment Research Council (NERC) wants to merge with the Agricultural and Food Research Council (AFRC), which in turn favours a single non-medical biological sciences research council. The Medical Research Council says the biology activities of the Science and Engineering Research Council (SERC) would be more sensibly placed with the three 'mission oriented' councils. The discussion has been conducted mainly before a House of Lords committee investigating the state of agriculture and food research in Britain.

Last week, in its evidence to the committee, the Advisory Board for the Research Councils (ABRC) announced that it is to set up a committee under Mr Richard Morris, the board's deputy chairman, to "review the disposition between research councils of responsibilities for research and training in the biological sciences and present arrangements for coordination" and "to recommend any changes . . . that are needed to improve efficiency and effectiveness".

On the question of its own role, ABRC chairman Sir David Phillips remains adamant that the board should not be given executive powers, but that it does have a wider role to play in the government's decision-making machinery. In particular, Phillips is "concerned" that his board should have closer links with the new Advisory Council on Science and Technology, notwithstanding the fact that he serves on both bodies.

On agriculture and food research, the board warns in its written submission to the committee that constraints on the science budget will result in AFRC being forced to reduce or completely eliminate some 20 high priority research programmes. Other important projects will remain consigned to the drawing board.

On commissioned research, the board is becoming "increasingly concerned by the significant reductions in research commissions from government departments, sometimes at relatively short notice". Effective forward planning by the research councils (principally NERC and AFRC) is "seriously undermined" and the financial and other costs of continual restructuring are a sizeable burden. The viability of research programmes part-funded from commissions and part from the science budget is put in question, the board says.

Simon Hadlington

Norwegian heavy water makes waves to the east

London

RELATIONS between Hungary and Romania, already tense over the Hungarian minority in Transylvania, worsened last week with reports in Hungarian newspapers that Romania has exported to Israel heavy water originally bought from Norway. But the issue seems to be a case of misdirected fire.

The Romanian denial, carried by the official press agency Agerpress, called the Hungarian reports "a violation of the elementary rules" of the media which "disinforms" the Hungarian and international public. What Romania overlooked was that the Hungarian reports echoed these already circulating in the Western press.

The first reference seems to have been by Professor Gary Milhollin, from the University of Wisconsin and a former consultant for the US Department of Defense, who described, in an article in *Le Monde*, Norway's unsuccessful efforts to inspect 20 tonnes of heavy water sold to Israel in 1959. He said that Israel had refused Norway access to its Dimona plant, but had said 8 tonnes of heavy water has been used up. He then noted that the 12 tonnes "they offer to show the inspectors" correspond "curiously enough" to the 12.5 tonnes sold by Norway to Romania in 1986. Milhollin was frankly speculating.

On 19 May, the Norwegian foreign ministry nevertheless announced an enquiry to find out whether Romania had indeed disposed of the heavy water, whose resale was forbidden without Oslo's consent. The heavy water was supplied to equip the Canadian-designed

Candu reactor on which construction began at Cernovoda in 1979. Of eastern European states, Romania alone has diplomatic relations with Israel.

By the time the Hungarian press took up the story on 24 May, Norway had demanded to know the location of the heavy water it had supplied. Two days later, a Romanian official in London said that the Cernovoda reactor (the first of five) was being filled with heavy water. Although Romanian officials claim that the reactor will be commissioned next year, the resident Canadian engineering manager, John Karger, says that even 1992 is a "very optimistic" date. Norway appears not to have known, when it sold the heavy water, that the Cernovoda plant was so far behind schedule.

Norway is already in difficulties over heavy water. Apart from the dispute with Israel over inspection of Dimona, it had to announce earlier this year that 15 tonnes of heavy water shipped to Frankfurt in 1983 had "gone missing".

Some Romanian sources suggest that the Romanian heavy water may have been sold to Israel to raise hard currency, and in the expectation that it would be replaced, by the time Cernovoda is commissioned, from the output of a small heavy-water plant at Turinu Sereniu.

Expert examination of the heavy water could provide an answer. The International Atomic Energy Agency in Vienna, saying that it has no statutory powers to inspect heavy water in Romania, says that it nevertheless has the necessary instruments.

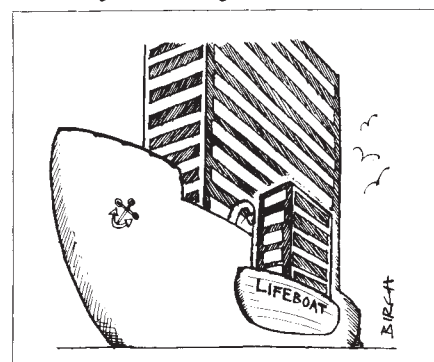
Vera Rich

Floating real estate in Tokyo Bay

Tokyo

LAND here can cost \$200,000 a square metre, so some bizarre ideas can make economic sense. Thus an Osaka company plans to build a small high-technology city centre on a 35,000-ton ship to be moored in Tokyo Bay. The influx of foreign securities firms and banks to Tokyo has so increased demand for office space that already astronomical land prices have been driven to even greater heights. Last month, the National Land Agency valued the Tokyo area's total land assets at 587 million million yen (\$4.7 million million), about twice the gross national product of Japan.

The Noah Project, to float a city sub-centre in the bay, was conceived by Temporary Center, an Osaka manpower firm, and is expected to be joined by Ishikawajima-Harima Heavy Industries and other shipbuilding companies. The ship will have office space for 700 people, a hotel accommodating 150 and a concert hall, and will



be equipped with advanced computer and satellite telecommunications facilities. A shuttle launch service will connect with the Tokyo waterfront.

Supporters of the scheme also point out that the ship would be safe from earthquakes and that, if a tsunami warning is issued, the "floating city" could head for the open sea.

David Swinbanks

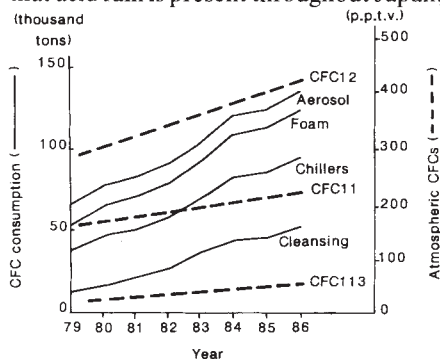
Counting the environmental cost of Japan's progress

Tokyo

JAPAN'S Environment Agency has passed a milestone in its history by making a self-conscious estimate of the effects of the phenomenon of Japan on the global environment. The agency's annual white paper, issued on 20 May, leaves few stones unturned. It deals with emissions of carbon dioxide, chlorofluorocarbons, nitrogen and sulphur oxides and acid rain, as well as tropical deforestation, trade in endangered species, and environmental problems in developing nations.

The view of Japan emerging is not especially flattering. Japan is a major source of chlorofluorocarbons, accounting for more than 10 per cent of world consumption. Much of the increase has been due to the use of chlorofluorocarbons in cleansing agents used by Japan's semiconductor industry.

The white paper also shows that Japan's emissions of carbon dioxide have grown much faster than those of western nations over the past two decades, now surpassing those of West Germany and Britain. And pH measurements by the agency indicate that acid rain is present throughout Japan,



Japan's annual consumption of CFCs by usage, and atmospheric levels of CFCs 11, 12 and 113. although the white paper depicts acid rain as predominantly a problem of Western Europe and North America. On the bright side, Japan's share of emissions of sulphur and nitrogen oxides has been held down by strong emission controls.

On trade, the report acknowledges that Japan is a major driving force behind the destruction of the world's tropical rainforests. Japan accounts for no less than 52 per cent of the world's imports of unprocessed tropical timber, the bulk of it from Malaysia.

Japan's influence as an importer of endangered species is openly discussed. The report says that imports of musk amount to 300 kg per year — entailing the destruction of more than 10,000 male musk deer a year. The deer is listed as an endangered species, but imports are allowed by classifying it as a "special reservation" species under the Washing-

ton Convention. But the agency says it expects the deer to be excluded from the reservation list from next year.

On remedies, the white paper urges that Japan should take "actions worthy of its position as a great economic power". In particular, it says that last year's Montreal Protocol for the protection of the ozone layer should be implemented "as soon as possible". (The Diet has already passed legislation and Japanese companies are rushing to develop alternative chemicals.)

In extending aid to developing nations, the agency says, Japan should follow the example of the United States, Britain and West Germany and include environmen-

tal impact assessments in planning for aid projects. Japan has so far left impact assessment to the recipients. In Malaysia, for example, Japanese aid has been used to build logging roads without consideration for the consequent deforestation.

Over and above all this, the white paper repeatedly stresses the need to "revolutionize the consciousness of every member of the Japanese nation". People must be inculcated from childhood with an understanding "of the relationship of the environment with daily life". And everyone should follow the "lifestyle of an Earthling".

These are strong words for an agency with little political clout in Japan. It remains to be seen if they are translated into actions. Most of the changes called for lie beyond the agency's domain.

David Swinbanks

Light at end of the tunnel for Germany's universities?

Munich

WEST Germany's science advisory council, the Wissenschaftsrat, has come to the defence of the university system with a series of recommendations for the decades ahead. Its report, *Recommendations on the perspectives of universities in the 1990s*, may be a valuable antidote to the gloom, repeatedly echoed at last month's Westdeutsche Rektorenkonferenz, that too many students and too little money over many years have diminished the once-lofty status of the university. The report is politically well-timed. Finance ministries in the *Länder*, which control university budgets, are seeking to cut corners as they face the impending tax reform.

But the Wissenschaftsrat recommendations carry a "high degree of obligation", according to Diether Breitenbach, a member of the council who is also Minister of Culture, Education and Science in the Saarland. The report was unanimously approved on 20 May, with the support not only of professors but also of the finance ministries.

The council is particularly distraught that good research is increasingly concentrated at the national laboratories (of the *Grossforschungsgemeinschaft*, the Max-Planck institutes and in industry rather than at universities. So the council recommends a sizeable increase of federal government support, which has fallen to 71 per cent of that in 1975, and which is expected to decline further. The federal government is also threatening to pay DM300 million less than originally planned for university construction projects in each of the next four years, but the *Länder* in this case say they are willing to pay their full half-shares of the original costs of planned projects.

The council fears that the trend for research to move out from universities will become even stronger, especially in applied fields such as electronics, which would further deepen the split between teaching and research, the union of which was the great triumph of the German university. The Wissenschaftsrat therefore urges that careful analysis should be given to proposals to found institutes outside universities to see whether their functions might not somehow be accomplished inside.

The Wissenschaftsrat says that university teaching must also be improved. It now takes students far too long to complete their studies (see *Nature* 333, 198; 1988). The average age for completing postdoctoral qualification (*Habilitation*) is 39.6 years, 10 years greater than in 1900.

The Wissenschaftsrat also advocates more competition between universities, a new departure in West Germany. Thus it says that universities should be free to choose their students in popular fields such as medicine, instead of having them assigned by a central office. Even the sacrosanct bureaucratic structure of the universities may be affected. Breitenbach says there is "no guarantee" that the current university personnel structure will be maintained.

The head of the Wissenschaftsrat committee that prepared the recommendations, philosopher Jürgen Mittelstrass of the University of Constance, says that a country like West Germany poor in natural resources cannot afford to cut back on education and research. The general-secretary of the Wissenschaftsrat, Peter Kreyenberg, hopes that the recommendations will help West Germany to avoid following the "horror example" of Britain.

Steven Dickman

Animal rights activists raise a storm in California

Berkeley

OVER the loud protests of animal rights groups, the California legislature has approved funding for a new \$12.5 million animal facility on the campus of the University of California (UC) at Berkeley. Construction is scheduled to begin in July.

The Northwest Animal Facility, in the planning stage for several years, has come under relentless attack by animal rights organizations, which see it as a sign of the university's continued commitment to animal research. The groups have drawn considerable public attention with campus demonstrations and press conferences. Because of these lobbying efforts, the state legislature's financing for the facility may come with strings attached: a legislative committee will decide by 15 June whether to ask UC to carry out a study of alternatives to animal research, or to place two animal rights activists on the Berkeley Animal Care and Use Committee (ACUC), the group responsible for enforcing the National Institutes of Health guidelines for animal care.

The university says it will try to honour the legislature's requests although they are not legally binding. It may balk, however, at placing animal rights activists on the committee, says ACUC chairman Rick Van Sluyters, because the pool of organizations from which the members would be chosen are all strongly opposed to animal research. As a gesture of good will, UC appointed an animal activist to the committee in 1985, but her presence was disruptive, says Van Sluyters, as she operated outside ACUC guidelines, trying to change rules the committee is authorized only to enforce. She resigned last November, criticizing the committee as a "rubber stamp" for researchers' interests.

Berkeley's five-year struggle with animal rights groups is a "case study in what can go wrong", says campus veterinarian Roy Henrickson, who directs the Office for Laboratory Animal Care.

When Henrickson took the job in 1984, he was faced with ageing, decentralized animal quarters that were difficult to regulate and maintain. The campus had lost its accreditation by the American Association for the Accreditation of Laboratory Animal Care (AAALAC), and a new, centralized animal facility seemed the logical way to regain it. That goal has been continually opposed by animal rights groups. Henrickson says he made the early mistake of dismissing the activists as "buffoons", not realizing that they would be accepted by the community as more credible than university officials.

Henrickson admits that animal care at Berkeley was flawed in the early 1980s,

variable in quality and an easy target for criticism. But the activists are doing their best to thwart the university's efforts to remedy the problems, he says. Henrickson believes Berkeley's problems are a harbinger of times to come for other research universities.

Just before legislative budget hearings on the new facility, the New-York-based

Fund For Animals launched a media campaign against Van Sluyters, a professor of optometry. His research on the development of the mammalian visual system was named "the most cruel, worthless animal research of the year," and leaflets were distributed on campus showing kittens with their eyes stitched shut. Van Sluyters countered by sending legislators a packet of letters of support from colleagues around the country, including Nobel laureates David Hubel and Torsten Wiesel, testifying to the quality of his work.

Marcia Barinaga

New cohesion for UK computer research

London

THE direction of Britain's government-financed research and training in information technology will in future be largely determined by a new panel of academics and industrialists in an effort to produce a coherent national strategy. The Department of Trade and Industry (DTI) and the Science and Engineering Research Council (SERC) has announced that a new 'advisory structure' is to be established, drawn from the academic and industrial sectors "in roughly equal numbers" and headed by a single advisory body reporting to both SERC and DTI on research strategy, resource allocation and individual applications for support (although both SERC and DTI will retain their normal procedures for awarding grants).

It is significant that a research council (which gets the bulk of its cash from the Department of Education and Science)

has agreed effectively to pool its entire budget for a substantial area of research with that from another government department. SERC's expenditure on information technology this year is expected to be £28 million on research and £14 million on education and training. DTI's expenditure on research will be £50 million.

The full structure of the advisory committees and the details of management arrangements have not yet been finalized, although it is envisaged that under the main committee, three principal subcommittees will concern themselves with devices, systems, architecture and systems engineering.

The hope is that the existence of a mechanism allowing a strategic overview will make it simpler for other government departments, such as the Ministry of Defence, to enter collaborative programmes.

Simon Hadlington

Hughes Institute to benefit minorities

Washington

THE Howard Hughes Medical Institute last week awarded grants totalling over \$30 million to undergraduate programmes in biology and related sciences at 44 colleges in the United States. These grants, awarded chiefly to liberal arts and historically black colleges, make up the largest amount ever given by a private organization for undergraduate education in the biological sciences.

The minimum grant is \$400,000. Four colleges receive over \$1 million, topped by Xavier University, a leading "historically black institution educating students in the sciences". The top four (ranked by amount received): Xavier University of Louisiana (\$1.8 million), Haverford College, Pennsylvania (\$1.2 million), Wesleyan University, Connecticut (\$1.2 million) and College of Wooster, Ohio (\$1.0 million).

The institute's president, Purnell W. Choppin, stressed the importance of ensuring "an adequate supply" of well-trained medical researchers in the future, and of increasing the number of people from minority groups entering biomedical

research careers. The institute's selection process also stressed the importance of interdisciplinary programmes.

The new undergraduate grants complement the Hughes Institute's plan to finance 300 doctoral fellowships per year in the biomedical sciences. Both these educational programmes are part of a campaign by the private research organization to fund science education in addition to basic research programmes.

Unveiled last October, this campaign will provide \$500 million to support science-related activities — primarily in education — over the next ten years.

Ironically, the news of the minority-targeted Hughes grants came just as a blue-ribbon panel chaired by former Presidents Jimmy Carter and Gerald Ford released a report announcing that the United States has "lost ground" in its efforts to achieve participation from minority groups in government, education and other institutions. Despite a growing pool of minority high-school graduates, according to the panel, a smaller percentage attend college today in the United States than in 1975.

Seth Shulman

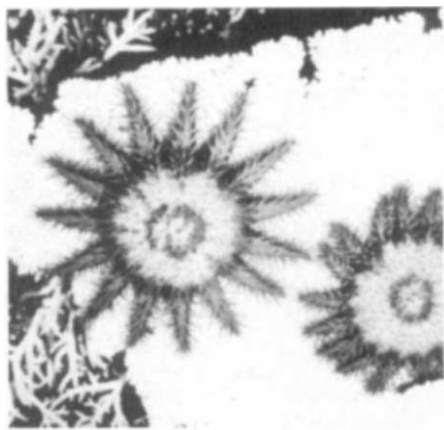
Thorny problem as Australia's coastline faces invasion

Sydney

REPORTS that Australian and Japanese companies are having second thoughts about building tourist facilities on the Great Barrier Reef have lent new urgency to the search for ways to control the Crown-of-Thorns starfish (*Acanthaster planci*).

The starfish continues to spread south, towards potential recreational sites, destroying large sections of reef. Yet Australian scientists are increasingly divided over the causes of the outbreak and what can be done to control it.

The first outbreak was recorded around Green Island, in the central third of the reef, in 1964. In 1979, during another outbreak at the same island, 3,000,000 star-



Rampant starfish.

fish were counted (the normal population density is about 6 starfish per square kilometre). By 1981, 90 per cent of the coral had gone and the starfish had moved to reefs further south.

Dr Bob Endine, associate professor at the University of Queensland, believes the starfish infestations are due to human intervention. "It is an ecocatastrophe of the first order" explained by over-collection of predators such as the Giant Triton. Reefs where predators have not been removed have remained "starfish-free", according to Endine.

Endine believes the only solution is to maintain a continual clearing programme and to prevent overfishing. He wants the government to ban the taking of molluscs and fish and wants crown-of-thorns starfish to be collected from the periphery of infested areas to prevent them from spreading. "A vigil must be kept", he says.

But many others are not convinced. Dr Peter Moran and Dr Roger Bradbury from the Australian Institute of Marine Science (AIMS), who are working on the still-obscure lifecycle of the starfish, believe the causes of the infestation will require better understanding of the reef and the relationship of the starfish to it.

As evidence that starfish control by clearing is ineffective, they cite a large-scale control programme in Japan where, over the past 20 years, approximately 13 million starfish have been removed from reefs in a designated region at a cost of A\$6 million. Starfish nevertheless remain, continuing widespread coral devastation.

Endine counters that control measures have previously been conducted in ways unlikely to succeed. He notes that the Japanese 'clearance' depended on a bounty system, meaning that only large starfish were picked up. In an Australian effort, "the army collected on only a section of a reef for two weeks", after which the starfish moved back.

AIMS and the Great Barrier Reef Park Marine Authority get A\$700,000 a year from the government over a four-year period, but senior project manager for research and maintenance of the Great Barrier Reef Marine Park Authority, Dr Leon Zann, is lobbying to have the funds

spread over a longer period of time.

"This is not a four-year problem, but will need decades of dedicated research." Zann also believes that the cause is yet to be known, saying that there is a need to know more about the life cycle of the pest and about the reef, which is why attention has turned to Fiji, Tonga and other Pacific regions where outbreaks have occurred. He notes that geological data show that starfish have been abundant in the past and that "this is not a modern phenomenon".

A major concern is the tourist dollar. With the coral as the drawcard on the reef, Japanese and Australian developers may decide not to invest in resort development because of fear of infestations.

The slow southward movement of the crown-of-thorns starfish is probably due to currents carrying larvae and the movement of adults in search of food. Endine believes that the fear for the tourist dollar is well founded. While moored submersible islands, floating hotels and man-made islands have been developed on the reef at more southerly sites to avoid the infestation, "the starfish are moving south too".

Tania Ewing

OECD report attacks national policies

Paris

THE Organisation for Economic Co-operation and Development (OECD), in a report* just published, says that member states are not paying enough attention to the potential uses of biotechnology in the control of environmental pollution. The report suggests that widespread public hostility towards genetic engineering could "cause considerable delays in the diffusion of many harmless and beneficial products and processes".

Awareness of this hostility, rather than empirical research, has, says the report, encouraged environment ministries to maintain a defensive position against hypothetical risks of the release of genetically altered micro-organisms, instead of exploiting the many ways in which biotechnology might help them to monitor and clean-up the environment. Positive examples — already being exploited in the United States and in France — include the use of biosensors as pollution monitoring devices and the development of improved waste treatment systems.

The lack of research to assess the real risks involved in the environmental release of genetically engineered micro-organisms is, claims the report, "rather disturbing". In the absence of criteria and methodologies for risk assessment, companies are often unwilling to embark on new avenues of research. This deadlock was found to be maintained by "lax regulations" in many member states which prefer to impose fines for waste emission, rather than to set up

efficient industrial treatment programmes. Also to blame is the lack of importance ascribed to risk assessment, which fails to attract sufficiently talented researchers.

More generally, traditional separation of biological and engineering faculties in universities, while gradually being changed serves as an obstacle to the interdisciplinary training of researchers needed to respond to a growing manpower shortage in industrial biotechnology. In France this common problem is compounded by the physical separation of the Grandes Ecoles, which train the most qualified engineering students, from the universities, where most research scientists are educated.

To improve private sector investment in what remains a high-risk, long-term research area, the report recommends international pooling of scientific information and the development of databases, regulatory mechanisms to ensure intellectual property and the freedom to publish findings (often an obstacle to university collaboration with industry), as well as better diffusion of information to maintain public confidence.

Given that more patent applications for commercially useful biotechnology inventions arise from basic research than from industrial, applied research, the report stresses that the biotechnology boom should not tempt governments to neglect the role of pure science.

Peter Coles

*Biotechnology and the changing role of government (OECD, Paris, 1988)

US *in vitro* fertilization in limbo according to OTA

Washington

INFERTILITY therapy has become big business in the United States. According to a new report* by the congressional Office of Technology Assessment (OTA), US spending on medical treatment to combat infertility amounted to \$1,000 million in 1987. But, for a variety of social and legal reasons, US research on treatments to improve reproductive success has not kept pace with the growing demand for the technology.

The number of infertile couples in the United States is estimated to have remained constant over the past 15 years at about 2.4 million, defined as those who have not conceived after 12 months of intercourse without contraception.

The report also says that about 20 per cent of all infertility is caused by sexually transmitted diseases.

If the percentage of infertile couples has not changed, the numbers seeking treatment have increased substantially. In 1968, there were 600,000 physical office visits for infertility services, but by 1984 the number had increased to 1.6 million.

Conventional medical and surgical treatment still seems to be the most common approach. But new non-coital technology, specifically *in vitro* fertilization (IVF) and gamete intrafallopian transfer (GIFT), are gaining acceptance, witnessed by the growing number of clinics providing these services. But the OTA points out that the federal government has virtually no role in supporting their development nor in monitoring their success.

Private organizations such as the American Fertility Society and the American College of Obstetricians and Gynecologists have supported some evaluations of IVF clinics, but OTA concludes that there is no way at present to tell whether IVF is an experimental or proven medical therapy.

Part of the explanation goes back to the 1974 National Research Award Act, which set up the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. That commission instituted local reviews of experiments involving human subjects by means of Institutional Review Boards, and also called for an ethical advisory board (EAB) within the Department of Health and Human Services (HHS). EAB was given authority to review grant applications for IVF research projects on a case-by-case basis, without which grants would be prohibited by HHS.

But EAB disappeared in 1980, as its charter came up for what should have been routine renewal at the end of the Carter administration and the then HHS

Secretary Patricia Harris let it expire. The board had become politically contentious because of its suggestion that certain types of research on IVF might proceed without review, which caused uproar from the potent US anti-abortion movement. The Reagan administration has not re-established EAB, although technically it is required to do so by HHS regulations.

Without a functioning board, the National Institutes of Health (NIH) cannot make grants for IVF research. Under normal circumstances, NIH reckon they would receive 100 grant applications a year for IVF-related research. NIH have tried several ways of breaking this legal logjam, including reviewing of IVF

research proposals before passing them on to HHS for approval by the non-existent EAB. So far, NIH have had no success.

But the Congress may be ready to intervene, in part because of the OTA report. The House of Representatives subcommittee on regulation and business opportunities of the Small Business Committee began hearings on 1 June on the commercialization of reproductive technology, emphasizing the need to offer at least some protection to consumers of IVF services.

As health insurance coverage for IVF is uneven at best, it is an expensive gamble. OTA estimates that, on average, IVF can provide a baby for one couple in ten at an average cost of some \$22,000.

Joseph Palca

*Infertility: Medical and Social Choices. US Congress, Office of Technology Assessment, Washington, DC 1988.

UK nuclear waste strategists still facing public suspicion

London

BRITAIN'S nuclear power industry continues to struggle in its efforts to gain public acceptance for a waste disposal strategy. Earlier this month, the local authority in Cumbria refused to grant permission for a test borehole at the site of the Sellafield nuclear complex, in spite of assurances from the operators, British Nuclear Fuels (BNFL), that the borehole would be used merely to obtain information about the site's geology as part of continuing investigations to assess the area's suitability for a deep depository for intermediate level waste. BNFL is still considering their response to the rejection.

The industry has ploughed tens of millions of pounds into its attempts to find a publicly acceptable means of disposing of low and intermediate level nuclear waste. Last year, in the face of severe local opposition, the Nuclear Industry Radioactive Waste Executive, Nirex, suddenly abandoned its investigations of four sites for shallow dumps for low-level waste in favour of deep depository for both low and intermediate level waste. Six months later, in November, Nirex produced a discussion document outlining the options available and inviting responses from interested parties. Nirex says it has sent out some 60,000 documents and has received around 800 responses, the deadline for which has been extended from the end of last month until mid-July.

The options under consideration are for a deep-mined cavity under land using conventional mining techniques; an off-shore sub-seabed cavity reached by tunnels on the coast; or an off-shore, sub-seabed cavity accessed from the sea's surface by a drilling platform or artificial island. Nirex

is thought to favour the first option, and experts have doubts about the feasibility of the third.

The government's Radioactive Waste Management Advisory Committee (RWMAC), whose chairman is Professor John Knill, of Imperial College and the chairman-elect of the Natural Environment Research Council, published its response to the Nirex document last week. The committee is not impressed with the government's current record on waste disposal, and has "viewed with continuing concern the several shifts in government waste disposal policy in recent years which have both caused delay in achieving suitable provision for disposal within an appropriate timescale and contributed to the public uncertainty over the practicability of any solution". The committee also wonders if Nirex should acquire in-house geological expertise instead of using the British Geological Survey, which as advisers to the government might be construed as having a conflict of interests.

Questions are also arising about the future of Nirex once the electricity supply industry is in private hands. Under government proposals, the Central Electricity Generating Board, which pays 42.5 per cent of Nirex's costs, is to be sold off as two companies, the larger of which will retain the nuclear capacity. Doubts are already being raised about the willingness or ability of the nuclear company to invest substantially in research. One suggestion is that the government would transfer responsibility for waste disposal to BNFL (which also pays 42.5 per cent of Nirex's costs), which is to remain in government hands and which produces most of the nation's nuclear waste. Simon Hadlington

Where science has gone wrong

SIR—The following are brief answers to some of the published objections (330, 308, 689–690; 1987 & 331, 129–130, 204, 384, 558; 1988) to our Commentary article (329, 595–598; 1987). The article was misinterpreted as an attack against philosophy, sociology and the human sciences, allegedly to preserve the prestige and privileges of the natural sciences. But the ‘human’ sciences are as scientific as the ‘natural’ sciences, and the allegation itself is also capable of being assessed objectively.

It is true that Popper and Feyerabend differ in many respects. But it is unfair to single out Feyerabend as worse than the others. For he too sets out to do what he sincerely believes is best for society. This he does by declaring “Anything goes”, a position that follows inescapably from the (popperian) premise that all observations are theory-laden. This premise is the most basic of all errors, and the writings of those who commit it are bound to be a hopeless muddle. If Popper, Kuhn and their followers do not like being called relativists, negativists and irrationalists, then they ought to repudiate this worst of all antitheses.

A frequent and passionate objection was to the dogmatic way in which we express some views. We never cease to be amazed by the frequency and vehemence of this objection — how can anyone seriously say with such passion anything at all if one does not have complete confidence in what one says? The reason why so many people are horrified by dogmatism is obviously because many instances in the past caused much harm. But in all these cases harm was done because what

people were dogmatic about was not true. If what one is dogmatic about is true, this dogmatism cannot possibly be harmful.

How can one be certain that one knows the truth? This is often difficult to answer. But we suggest that this is the very question that every professional (not only scientists and philosophers but also historians, physicians, journalists, police officers and so on) ought to be trying to answer, instead of denying the very existence of truth. If one does not do so, this conduct should be seen for what it really is: a breach of professional duty. The question is whether on the available evidence a hypothesis has been refuted, or verified (and how accurately), or is still open to further investigation. In the latter case one has to be sceptical, in the former dogmatic.

It would be irrational, and probably harmful, not to be dogmatic in one's advocacy of a proven truth. A good example may illustrate this point: Consider the proposition that “when ingested, 0.3 grams of sodium cyanide will rapidly kill a person”. We believe that this is true, we are absolutely certain about it, and we are dogmatic about it. We invite those critics who think that our dogmatism is wrong and dangerous to refute it.

It was also objected that science attained its present level without much self-analysis, and that most major breakthroughs were apparently made by accident, chance, serendipity or at best with Popper's prescription of ‘trial and error’, ‘hit and miss’, ‘conjecture and refutation’ methodology. But what is this, if not the methodology of the Stone Age? To those who wish to move forward and

make progress, we would recommend much more serious self-analysis and forward planning. The last objection is only partly correct, and for a fuller picture we recall the old saying “luck favours the prepared” and Rutherford's famous dictum “an ounce of thought is worth a ton of equipment”.

Nature's hostility towards ‘creation science’ was recently rebuked in these terms: “Creationism is no more untenable a concept than the postulate of modern physics that matter can be created from nothing”¹. While we have no dogmatic opinion as to how the Universe originated, we are fairly certain that ‘matter can be created from nothing’ is not a true postulate of physics. In fact the epistemological antitheses to some extent owe their resurgence in the twentieth century to this and other alleged ‘postulates of modern physics’ of a similar nature. But these ‘postulates’ do not follow from any rational study of the evidence, and, contrary to what was suggested, there seem to be a few things wrong with modern physics. The sooner these errors are corrected, the sooner people will be able to see through the fallacies of the antitheses, of creationism, the paranormal and so on. It also appears that these errors in physics were the result of the physicists' poor grasp of the scientific method. For these errors are the ‘solutions’ of certain scientific problems arrived at by applying not the scientific method as such but instead the theological method of the Christian religion².

T. THEOCHARIS
M. PSIMOPOULOS

*The Blackett Laboratory
Imperial College,
London SW7 2BZ, UK*

1. Chelvam, R. T. *Nature* 331, 10 (1988).
2. Stannard, R. *The Times*, 3 December 1983.

Access to journals

SIR—The letter from Edward D. Goldberg (*Nature* 332, 10; 1988) emphasizes an important problem in the area of scientific communication among its practitioners. Although international agencies may be able to help alleviate the shortage of scientific journals in certain countries, they may not be able to do so in the near future. A voluntary effort by concerned scientists in North America and Western Europe might perhaps prove to be more useful.

It is common for working scientists to subscribe to one or more scientific journals, one general and one specialized. It is far too common for these journals to be thrown away after one year on the shelf. Certainly, every summer, in the medical school building at this university, corridors are lined with discarded copies of *Science*, *New England Journal of Medicine* and *Cancer Research*, to name a few. It should be possible for us to ‘adopt’ an institution with limited journal holdings to

which most of the year-old scientific journals from a given university might be mailed. One-year-old scientific journals are better than none at all.

SALEEM JAHANGEER

*Institute for Disease Prevention,
George Washington University Medical
Center,
2300 Eye Street, NW,
Washington, DC 20037, USA*

Risks of AIDS

SIR—In a news item “AIDS and public opinion in France” (*Nature* 332, 295; 1988), we were invited to wonder that “even those with a baccalaureate . . . felt that human immunodeficiency virus (HIV) could be transmitted by mosquitoes [or] dental instruments . . .”. The probability of transmission by sexual contact is of the order of one in 10. What is the working definition of “could be transmitted” — one in 10⁴? 10⁷? Surely the answer must depend on the size of the

population facing a minor exposure threat and subjective factors. Where are the data to verify the null hypothesis to such accuracy for these not implausible transmission routes? Would the designer of the French questionnaire really put a contaminated dental instrument in his mouth?

It is as irresponsible here as in nuclear power safety for scientists to assert that events of small or indeterminate probability cannot happen. The larger issue is the level of conflict between the civil and human rights of AIDS sufferers and the rights of the complement to avoid exposure. It serves no useful purpose to pretend the answer is zero. I hope it is an underestimate of human goodwill and tolerance that unless we are convinced the answer is zero, the rights and needs of both groups cannot be substantially accommodated.

MICHAEL H. FREEDMAN
*Department of Mathematics,
University of California, San Diego,
La Jolla, California 92093-0112, USA*

Towards a better control over science

SIR—The desire to understand the world has today been overtaken by the necessity of exploiting it. Scientific results and advances are for the most part produced within institutions whose ultimate aims are technological. The direction in which research, whether 'fundamental' or 'applied', proceeds is governed by economic, social, health-related or military considerations.

This orientation cannot be ignored by research workers and society has the right to pass judgement on it. With its reductionist viewpoint and disregard for all other forms of truth and knowledge, science puts itself in conflict with nature, culture and individuals.

Thus, unless science can be brought under control, it represents a serious risk to the environment, to mankind and to individuals. Scientific development continues, however, and promotes its own acceleration, with the naive assent of a society whose only vision of the future is expressed in terms of technological artefacts, although the identification of scientific achievement with progress, or even happiness, is largely imaginary. The increasing rate of scientific output is causing a qualitative change in the dependence of the individual on science. This is obvious for the practical aspects of life, which are constantly being altered by technology, but also holds for its more private aspects. The concepts of subjectivity, privacy and secrecy are being vigorously attacked by scientific disciplines which probe deeper and deeper and, while they are unable to understand everything, claim to explain everything. In the name of scientific truth, life is reduced to its measurable aspects. The increasingly narrow specialization of scientists encourages their shortsightedness as to their role in society and creates insurmountable barriers between the different scientific disciplines.

It is certainly difficult to turn back from technological advances resulting from scientific endeavour, which themselves create new demands in an industrial spiral which neither scientists nor consumers can control. We believe that clear thinking is more important than efficiency and the direction of research more important than the speed at which it is accomplished. We believe that reflection should precede any scientific project rather than following any innovation made. We believe this reflection has to be philosophical rather than technical in nature, and that it should take place across interdisciplinary barriers and include the whole of society.

J. ARSAC (Informatique, Univ. Paris VI); M. BOUNIAS (Biochimie, Univ. Avignon); M. CASSE (Astrophysique, CEA Saclay); J-P.

DELEAGE (Physique, Univ. Paris VII); A. GSPONER (Physique des hautes énergies, ISRI, Genève); A. JACQUARD (Génétique, INED, Paris); J-M. LEVY-LEBLOND (Physique Théorique, Univ. Nice); J-M. MEYER (Embryologie, Univ. Strasbourg); F. B. MICHEL (Pneumologie, Univ. Montpellier); J. PANIJEL (Immunologie, CNRS-Pasteur); B. PRUM (Statistique médicale, Univ. Paris V); J-P. RENARD (Embryologie, INRA-Pasteur); J-C. SALOMON (Cancérologie, CNRS, Villejuif); J-P. SCHEIDECKER (Astronomie, CNRS, Nice); J-P. SHAPIRA (Physique nucléaire, Orsay); M. SINTZOFF (Informatique, Univ. Louvain); J. TESTART* (Biologie, INSERM, Clamart).

*Corresponding author: J. Testart, INSERM U. 187, Hôpital Antoine Bécclère (92141) CLAMART, France.

Ozone depletion

SIR—Some of the views in your recent leading article on ozone¹ cannot be allowed to pass unchallenged. Those working in the field will smile wryly at the thought that financial support has been generous. Some satellite instruments were designed under cost constraints that precluded planning for extended operation. Long-term monitoring activities have been despised as 'routine', and support for them severely cut back. Problems vary from country to country. In the United States, it seems easier to get support for building instruments than for providing the resources to make prompt and good use of the data acquired. In Britain, lack of capital (and over-rigid accounting rules) has severely limited the capacity to build or deploy suitable instruments. World-wide, many major field and laboratory studies could not have been undertaken without substantial support from the Chemical Manufacturers' Association.

Insufficient funds also explain why the arguments used in discussing the environmental effects of ozone depletion are so often emotive. Recommendations have been made at intervals since 1974 for comprehensive long-term research into the effects on man, animals, plants and materials of increased fluxes of ultraviolet radiation. They have yet to be acted upon.

The most pressing question, however, is this — is current understanding of the processes that control the abundance of ozone sufficient to justify existing policies on halocarbon emissions? There is no longer any doubt that halocarbons are primarily responsible for the effects observed in Antarctica. The Ozone Trends Panel has found definite evidence outside Antarctica for ozone depletion over two decades, overlain by effects of the solar cycle. The observed cyclic change of ozone is simulated fairly well by models. However, when allowance is made for increase of tropospheric ozone², these models significantly underestimate the underlying trend in stratospheric ozone. There are no grounds for compla-

cency — the assumption in the Montreal Protocol that the halogen content of the atmosphere can be allowed to increase is no longer tenable. The protocol must be swiftly ratified, and strengthened.

J.C. FARMAN

*British Antarctic Survey,
High Cross,
Madingley Road,
Cambridge CB3 0ET, UK*

1. *Nature* 332, 291 (1988).

2. Penkett, S.A. *Nature* 332, 204–205 (1988).

Stations in space

SIR—The anonymous writer of your editorial "Stations in space" (*Nature* 332, 292; 1988) seems not to have thought very deeply about the future of space exploration. The most important 'purpose' of the space station is the experience that will be gained during its construction. All other suggested purposes are, to some extent, icing on the cake. The future of space exploration, including that of space science carried out from Earth orbit, will largely depend on the ability to manufacture large structures in space, and the building of the space station will pioneer the necessary technology.

It should be self-evident that the continued manned (or even large-scale unmanned) exploration of the Solar System will depend on the assembly of large spacecraft in Earth orbit. The experience to be gained by the construction of the space station will be an essential prerequisite to these ventures. Moreover, while your writer is quite right about the immense value to astronomy of the space telescope, this instrument should not be seen as the culmination of space astronomy. It is certainly possible, in principle, to build larger instruments, but, again, these will have to be assembled in space and the above argument is just as valid. Important though the space telescope is in the short term, history is likely to show that the ability to construct large space structures (telescopes, space ships and so on) will be of more lasting scientific value. Make no mistake, this technology, carried to its logical conclusion, will enable us to send instruments to the stars.

Finally, large-scale investment in space is important because it provides business for high-technology companies which would otherwise be producing weapons. This is important because only when such alternative business is found will disarmament become acceptable to the US military-industrial complex. In fact, given the magnitude of US military spending, the argument that the federal deficit makes the space station an irresponsible undertaking looks a little silly.

IAN CRAWFORD

*University of London Observatory,
Mill Hill Park,
London NW7 2QS, UK*

Looking for the unseen Universe

Physicists may soon be able to detect particles which cosmologists tell them constitute most of the Universe. But the experiments are both speculative and expensive.

THE cooperative study of particle physics and cosmology was established as a worthwhile scientific endeavour by B. W. Lee and S. Weinberg in 1973. In an early and influential paper, they showed that simple arguments in cosmology could constrain possible values of the neutrino mass which, according to some then-current theoretical preferences might not be zero. Neutrinos are produced in the early Universe in accurately predictable numbers and, if they have mass, their contribution to the density of the Universe today can easily be calculated. But astronomers know unequivocally that visible matter in the Universe today cannot be as little as, say, one thousandth of the mass of the unknown invisible component, which sets an upper limit for the neutrino mass.

This illustrates the general style of argument in cosmology and particle physics, although the chains of reasoning have become more intricate and more hedged about. For example, if a neutrino of a particular mass is ruled out by Lee and Weinberg's calculations, then it can be saved by assuming it is unstable, and decays into massless particles. But then one must worry about what the decay products do to standard cosmological evolution. They may heat the Universe or distort the microwave background or upset the synthesis of helium, deuterium and lithium in the first three minutes of the Big Bang.

If cosmologists are ingenious at thinking up arguments to constrain what elementary particles can do, particle physicists are just as ingenious at evading the constraints. The neutrino, for example, can be made to decay into something invisible and innocuous whose presence the most assiduous of cosmologists cannot detect. If there was a single theme that emerged from a recent gathering of astronomers and physicists*, it was the need for hard facts, from cosmology or experimental physics, to set boundaries on the imagination of theorists. Massive neutrinos were mentioned often, and enlisted in a variety of causes, but what mass they have, if any, is little better known now than in 1973.

Over the years, the relative roles of particle physics and cosmology have changed. At first, simple cosmological models were used to test and provide limits on exotic theories of particle physics. But cosmologists have begun to invent more sophisticated evolutionary models

dependent on particles which, it is hoped, physicists will one day be able to supply. This has given participants on both sides of the enterprise a bad name.

There is still no satisfactory theory of the formation of galaxies. Observers have catalogued galaxy positions and redshifts in increasing numbers, but their distribution in space still defies accurate description. Galaxies congregate on filaments and knots, and there are large regions of space devoid of them. But this empirical picture has not been made quantitative. And there is no consensus on whether the empty regions — the voids — are empty of matter or only of matter in luminous form. In short, astronomers cannot say how visible matter is distributed in the Universe and do not know whether the invisible matter is distributed in the same way as galaxies, or more uniformly.

This leaves cosmologists with some difficult problems to solve and a good deal of latitude in which to solve them, which encourages speculation. The most serious obstacle that galaxy formation theories must negotiate is the absence of measurable temperature fluctuations in the microwave background. The Universe at 100,000 years of age have been remarkably free from irregularities, but now, when it is ten thousand million years old, there are superclusters and voids. Finding ways for these enormous structures to grow from almost nothing in the permitted interval is difficult; cosmologists suppose that the chief agent is whatever material constitutes the dark matter, and that the visible and familiar stuff of galaxies merely falls into place later.

A trap into which cosmologists can fall is to develop a theory around some putative elementary particle, whose properties they particularly like, to describe a large-scale distribution of matter so as to explain the presence and arrangement of galaxies which make up a dynamically insignificant fraction of the modern Universe.

There is now a range of theories of galaxy formation sharing some uncontroversial elements but whose finer differences cannot be resolved for lack of data. The good news in Bologna was twofold: cosmologists are anxious to find independent tests of their ideas, and physicists have some hope of providing them.

Massive neutrinos have been supplanted as the most popular dark matter candidates by axions. (Neutrinos cannot explain the small-scale structure of galaxy cluster-

ing.) Axions, a concomitant of the Peccei-Quinn explanation of CP conservation in strong interactions, are much less massive but much more numerous than cosmological neutrinos, and have lower velocities. These persuade cosmologists that it has a much better chance of playing the dynamical role required of dark matter constituents. The greatest drawback of the axion, though, is that there is not one piece of experimental evidence that it exists.

An experiment now in progress at Brookhaven National Laboratory, New York, may provide that evidence, or discount the axion as a dark matter candidate. In the presence of virtual photons provided by a strong magnetic field, axions can convert into microwave photons at a potentially measurable rate, if indeed we are bathed in a sea of cosmological axions.

A tunable microwave cavity inside a large superconducting magnet thus becomes a cosmological axion detector. The frequency of the detectable microwave photons is directly proportional to the axion's mass, so by tuning the cavity, axions in a likely mass range can be sought out. There is a variety of other constraints on the axion mass, from laboratory experiments as well as from astrophysics, making a mass from 10^{-3} to 10^{-5} eV the only possibility.

Unfortunately, the sensitivity of the experiment is about an order of magnitude less than what is needed, and the lack of detection so far is as expected. But the sensitivity of the detector increases with the volume and the square of the magnetic field, and it seems reasonable to imagine that a thousandfold increase could be obtained. This would allow detection or dismissal of the axion hypothesis with equal firmness.

But the problem now, in blunt terms, is to decide how much it is worth spending to test one admittedly speculative theory of the structure of the Universe. The present experiment used an old magnet built some years ago at Brookhaven, but a bigger detector would require a custom-built and consequently expensive device. And those who spent a couple of years designing and performing this prototype experiment may prefer not to devote the many years needed for the bigger version.

David Lindley

* Third ESO-CERN meeting on astronomy, cosmology and fundamental physics, 16-20 May, Bologna.

Molecular recognition

Viral receptors and drug design

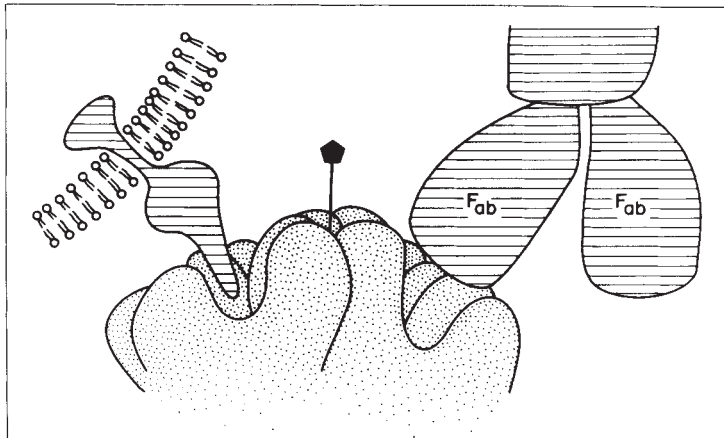
Michael G. Rossmann

A CRITICAL event in the life-cycle of most animal viruses is their attachment to specific receptors on the cell surfaces of hosts. This sets the stage for penetration into the cell, either by fusion of the viral membrane with host cell membrane or the creation of endosomal vesicles. Eventually the virus, or virus core, 'uncoats' and proceeds to subjugate the host cell's machinery for synthesis of viral components that are able to recognize each other in the assembly of new viral particles. The latter events give rise to viral pathogenesis. But it is the attachment step which is often the main determinant in host specificity and tissue tropism, the consequence of a molecular recognition by a viral surface component for a host cell receptor. The first three-dimensional view of such a process is given by Weis *et al.* on page 426 of this issue¹ in their study of influenza virus.

Influenza viruses are enveloped by a host cell-derived lipid membrane which is penetrated by numerous copies of two distinct types of virally coded molecules: trimeric haemagglutinin and tetrameric neuraminidase. The former is responsible for cell recognition. The relative molecular mass of the haemagglutinin trimer is around 250,000 and its structure² consists of an 85-Å-long protein fibre and a large globular head. Four non-overlapping epitopes that bind neutralizing antibodies have been mapped onto the surface of each monomer³. Mutations that cause variation of amino acids in these sites give rise to antigenic drift and, if such changes occur in all four epitopes, then there is antigenic shift, leaving large proportions of the population unprotected against influenza virus infection. There is a small depression in the haemagglutinin surface between three of the four neutralizing epitopes. Residues lining this depression or pocket are highly conserved in comparison with the hypervariable antibody binding sites. This surface feature has now been shown to be the binding site of sialyllactose, a trisaccharide-receptor analogue.

Viruses that infect vertebrates are under constant surveillance by the host's immune system, producing pressure for

the virus to evolve ever-changing surface properties to escape neutralization by antibodies. This requirement is, however, contradictory to the need of the virus to conserve receptor specificity. The surface of rhinoviruses⁴, as well as other picornaviruses, contains large 'canyons'. It has



Depressions on viral surfaces suggest a strategy for evasion of immune surveillance. The shape of the putative receptor-binding regions sterically hinders antibody (top right) recognition of residues at the base of the site, while still allowing recognition and binding by a smaller cellular receptor (top left). This would allow for receptor specificity while at the same time permitting evolution of new surface properties that alter the viral antigenicity. Although the diagram pertains to a study of human rhinovirus, where the receptor attachment site is a 'canyon' circulating about each of the 12 5-fold axes (one of these is designated here by a black pentagon), similar principles are equally applicable to the haemagglutinin spikes of influenza virus and, no doubt, to many other viruses. A possible antiviral strategy would be to design small compounds that mimic the cell receptor and thus preferentially bind to the virus, thereby blocking attachment.

been suggested that these depressions are required to protect the receptor recognition site on the virus from antibody attachment (see figure). That this canyon is, indeed, the site of receptor attachment has been substantiated by using site-specific mutagenesis of residues lining the canyon⁵. A similar situation clearly pertains to the receptor attachment site of influenza virus and is also relevant to substrate binding on the influenza neuraminidase molecule⁶. The same principle is likely to be extended to many other viruses. The conserved receptor-attachment regions on viral surfaces might thus be the Achilles heel for attack by antiviral compounds that are small enough (in contrast to antibodies) to enter the surface depressions or canyons and can bind independent of strain or serotype.

Many cell surfaces — erythrocytes, for example — have carbohydrate moieties covering large parts of their surfaces. Sialic acids (there are more than 20 naturally occurring derivatives of neuraminic acid, which form the family of

compounds named sialic acids) are usually found in the terminal position of oligosaccharide chains. These sugars are frequently used as receptors for hormones and other cellular functions, but are also perverted by viruses, such as influenza, as well as by some toxins. Rogers *et al.*⁷ had shown that a mutation of the amino-acid residue leucine 226 to glutamine of influenza virus haemagglutinin alters the specificity of haemagglutinin from oligosaccharides, in which the terminal neuraminic acid derivative is linked by an $\alpha(2,6)$ bond with lactose, to a specificity for an $\alpha(2,3)$ link with lactose. They therefore concluded that residue 226, situated in the conserved depression, would be associated with receptor recognition.

Weis *et al.* in this issue¹ now report on the detailed atomic interactions made by sialic acid, linked both by $\alpha(2,6)$ and $\alpha(2,3)$ to lactose, on binding to native and mutated (Leu 226 $\rightarrow$ Q) haemagglutinin. They find that nearly all the lactose moiety is disordered but that the sialic acid component is clearly visible. Although they are not able fully to understand the cause of the altered specificity of the mutant, knowledge of the mode of binding may permit the design of drugs which are able to compete with the cell receptor and thus block viral attachment to host cells.

Until recently, the only satisfactory defence that could be mounted against

viral infection has been the harnessing of the host's immune system by prior exposure to noninfectious mimics of the real pathogen. This strategy is unlikely to work where there are numerous serotypes, as is the case for common cold (rhino) viruses or, to a lesser extent, for influenza viruses. However, in the past few years there has been an increasingly intensive effort to find antiviral agents targeted against specific and essential steps in the viral life-cycle. Indeed, antiviral agents are the principal hope in controlling human immunodeficiency virus, the AIDS pathogen.

Nevertheless, to date only very few antiviral agents are available for clinical use. The difficulty in their design is that they must be exceedingly specific to viral products for otherwise they will interfere with normal metabolism of the cell and, hence, exhibit toxicity. For instance, substances that block viral attachment should not inhibit attachment of those hormones that use the same cellular receptors. The best targets are, therefore, those which

pertain explicitly to those viral functions for which there are no analogous processes in the host cells⁸.

It is also necessary to establish routes of delivery which permit an accumulation of appropriate drug concentration at the anticipated site of infection. These are by no means the only difficulties in drug design, for it should not be thought that, given a lead structure such as sialyllactose, that it is then straightforward to design compounds that bind better to haemagglutinin. However, Weis *et al.* make some interesting intuitive suggestions. But perhaps it may not be too many years before proliferating computer programs

have been taught enough structural chemistry to enable them to outdo their human teachers. □

1. Weis, W. *et al.* *Nature* **333**, 426–431 (1988).
2. Wilson, I.A., Skehel, J.J. & Wiley, D.C. *Nature* **289**, 366–373 (1981).
3. Wiley, D.C., Wilson, I.A. & Skehel, J.J. *Nature* **289**, 373–378 (1981).
4. Rossmann, M.G. *et al.* *Nature* **317**, 145–153 (1985).
5. Colonno, R.J. *et al.* *Proc. natn. Acad. Sci. U.S.A.* (in the press).
6. Varghese, J.N., Laver, W.G. & Colman, P.M. *Nature* **303**, 35–40 (1983).
7. Rogers, G.N. *et al.* *Nature* **304**, 76–78 (1983).
8. Rossmann, M.G. *Proc. natn. Acad. Sci. U.S.A.* (in the press).

Michael G. Rossmann is Hanley Professor of Biological Sciences at Purdue University, West Lafayette, Indiana 47907, USA.

Conservation biology

Urban extinction of birds

Jared M. Diamond

HABITAT destruction is a significant factor in the worldwide crisis of species extinctions. But preserving a particular habitat does not guarantee preservation of species, because populations isolated in dispersed habitat patches are prone to extinction — even when the patches are as large as the largest national parks (Newmark, W. D. *Nature* **325**, 430–432; 1987). Thus we need to understand the causes and rates of extinctions, and to identify the most vulnerable species, in fragmented 'islands' of habitat. M.E. Soulé *et al.* (*Conserv. Biol.* **2**, 75–92; 1988) have

now analysed extinctions in a familiar but little-studied setting: the scraps of natural habitat left over by urban sprawl.

The present site of San Diego (see map) was formerly a mediterranean scrub habitat called chaparral. As flatter terrain was cleared for development, chaparral — and, with it, many native species — became restricted to steep canyons. But escalating land values and modern earth-moving equipment have led to development even there. Soulé *et al.* took a census of species in 37 chaparral-containing canyons ranging in area from 0.4 to 103

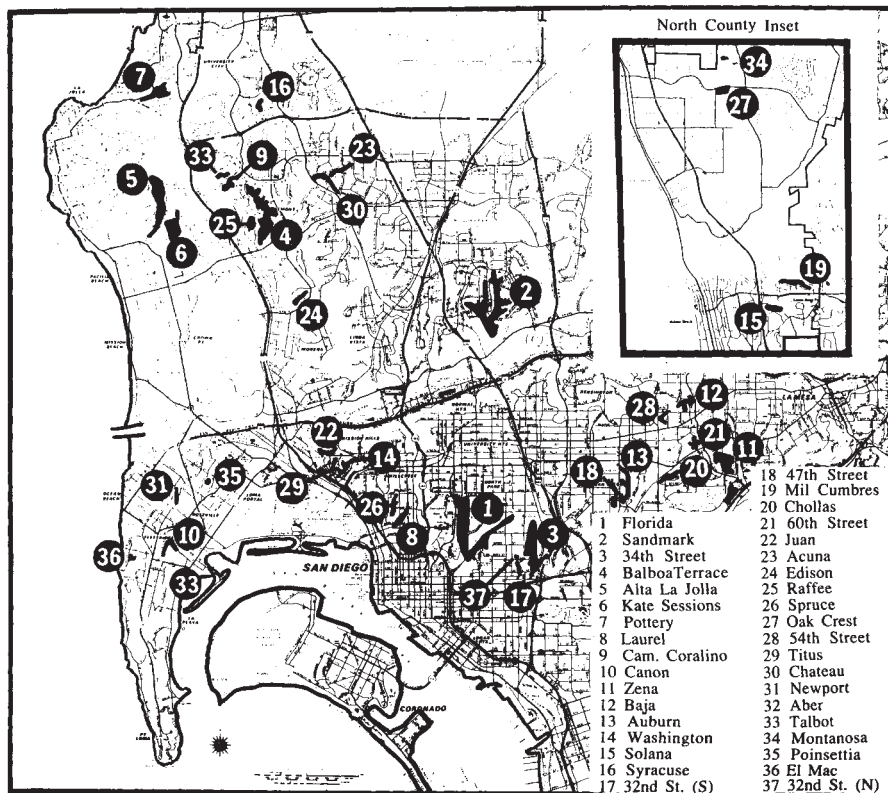


Top, rufous-sided towhee and bottom, roadrunner, a ground-dwelling cuckoo confined to chaparral, both of which gradually disappear from isolated canyons. (Photos: top, James Karo; bottom, Terry Root.)

hectares, and in 'age' (length of isolation by development from other chaparral areas, as determined by aerial photographs and building department records) from 2 to 86 years. They analysed survival of those seven species of native birds for whom chaparral is essential for breeding in this region.

All the canyons must originally have contained all or most of these 7 species, as similar-sized areas of nearby unfragmented chaparral, plus even the smallest canyons isolated from other chaparral areas for less than 15 years, contain 5–7 species. Many of the canyons have now lost some or all of the species that were there originally. Soulé *et al.* identified four variables that collectively account for almost all the variation in numbers of canyon species: total area; the area occupied by chaparral (both of which determine population sizes of chaparral species); duration of isolation, which determines the time available for populations to become extinct; and, surprisingly, coyotes and foxes. Large, recently isolated canyons with much chaparral, coyotes and no foxes have the most species. Solano Drive (7.6 hectares, most of it still chaparral, isolated for only 11 years, coyotes but no foxes), for example, still has all 7 species, whereas Spruce Canyon (4.3 hectares, hardly any of it chaparral, isolated for 86 years, foxes but no coyotes) has no surviving chaparral bird species.

Demographic fluctuations, inbreeding and loss of heterozygosity in small isolated populations, combined with the effects of predators and climatic variation, probably caused the extinctions. Original estimated population sizes of each species were only



Map of San Diego and environs, showing locations of isolated canyons with surviving chaparral habitat. (Photo courtesy of Michael Soulé.)

1–10 individuals for 2-hectare canyons and 1–100 individuals for 20-hectare canyons. A few decades of isolation eliminated about half these small populations, and 70 or more years of isolation eliminated almost all populations.

It seemed paradoxical at first that survival of birds should be promoted by coyotes, which are large predators that eat birds on occasion. But coyotes prey more heavily on medium-sized predators such as foxes, cats, raccoons, opossums and skunks, which in turn prey more heavily on birds than do coyotes. Hence local extinction of coyotes leads to population explosions of medium-sized predators and thence to bird extinctions. This phenomenon, which Soulé *et al.* term "mesopredator release", is likely to be of widespread importance.

Some chaparral species are much more prone to extinction than others. The characteristic sequence in which species tend to disappear after a canyon is isolated begins with black-tailed gnatcatcher and proceeds through roadrunner, California quail, California thrasher, rufous-sided towhee and Bewick's wren to wren tit as the last holdout (see figure). This sequence is close to that of the species' abundances in undisturbed chaparral (0.25, 0.02, 0.96, 1.10, 1.29, 1.75 and 2.5 pairs per hectare, respectively), as populations consisting of fewer individuals are likely to become extinct sooner. A second factor affecting this sequence is body size: big species (such as the large but rare roadrunner) can better survive short crises in availability of resources and in weather conditions. These two predictors, abundance and body size, account for almost all the variation in canyon occupancy.

Another surprising result of Soulé *et al.* is that species survival was not affected by distance from colonization sources, the biogeographical variable usually second in importance only to area. However, most of the canyons now lie 100–2,900 metres from the nearest canyon containing the common chaparral bird species, but those species are notoriously sedentary and refuse to cross strips of non-chaparral habitat wider than 50–100 metres. Thus, once development isolates a canyon by more than 100 metres, chaparral birds cannot immigrate, so Soulé *et al.* suggest that the simplest way to reverse extinctions in canyons would be to connect them to other chaparral areas. Because chaparral birds are observed in strips as narrow as 1–10 metres, even strips left along the edges of roads and under power lines could prevent extinctions. Reintroductions of chaparral species into those canyons that have already lost them could reanimate the San Diego landscape. □

Jared M. Diamond is a professor of Physiology at the University of California Medical School, Los Angeles, California 90024, USA.

Planetary satellites

Taking a dim view of Nereid

J. Veverka

NEPTUNE has two known satellites. One, Triton, gets all the attention because of its retrograde and probably unstable orbit, its atmosphere of nitrogen and methane, and the possibility that it has an ocean of frigid liquid nitrogen¹. The other is neglected Nereid, a faint satellite discovered in 1949 by G. P. Kuiper, about which little is known. Nereid orbits Neptune once every 360 days in an inclined, highly eccentric path² which takes it from 54 R_N out to 400 R_N (R_N is Neptune's radius, about 25,000 km). The unusual orbit fostered the belief that the satellite is a captured object, although some suggest³ that Nereid was once a normal moon that was perturbed into its current elongated path by a passing body. If Nereid is a captured object, it is almost certainly not a comet. Its diameter must exceed 200 km, too large for a comet, but comparable to that of the unusual object 2060 Charon, which orbits between Saturn and Uranus⁴. Whatever Nereid's origin, it has seemed unlikely that its spin period has been synchronized by tides to its orbital period of 360 days⁵. Now, careful photometric observations by Schaefer and Schaefer, reported on page 436 of this issue⁶, show that the spin period is less than 1 day, and also suggest that Nereid must be a very odd satellite that deserves considerable attention.

Schaefer and Schaefer find that Nereid could have an unusual colour and definitely displays large light variations of about 1.5 mag (a factor of 4 in brightness) with a probable period of between 8 and 24 hours. Clearly, it is not in a synchronous spin state, but more observations are needed to refine the period and obtain data on the position of the pole. The data reported in this issue contain no direct information on Nereid's albedo; hence the satellite's size remains undetermined. Because of its faintness, Nereid's infrared colours remain to be measured; it has not been possible to apply the powerful 'JHK technique' (using the J-, H- and K-bands of the infrared spectrum) developed by Cruikshank and Hartmann^{7,8} to distinguish icy objects from rocky ones and obtain useful estimates of albedos for distant bodies.

Whatever its size, Nereid must be an unusual object. If the large light variation arises from an irregular shape, the satellite's longest axis must be four times bigger than the shortest dimension — making it much more elongated than any other satellite. Otherwise, some of the light variation must be caused by spots on the surface: Saturn's satellite Iapetus is spherical, for example, but still has an

intensity variation of 1.8 mag. But this explanation also makes Nereid, like Iapetus, a bizarre object with highly contrasting albedo markings, indicating large-scale surface exposures of segregated and very distinct materials (such as ice and carbonaceous rock).

As already noted, Nereid's albedo is still unknown. Average diameters of known objects span the wide range from about 210 km for an albedo of 1.0 (Enceladus) to about 1,500 km for an albedo of 0.02 (darkest asteroids). If Nereid is much larger than 300 km, then it is unlikely to be very irregular. All satellites larger than Saturn's Hyperion (average diameter about 250 km) have been found to be almost spherical⁹.

So what is weird about Nereid: a very irregular shape or highly contrasting markings? Fortunately, the answer will not be long in coming. Voyager 2 is on its way to Neptune and in late August 1989 will approach Nereid within 4.7 million km, close enough for the high-resolution camera to resolve details of about 50 km. Even if Nereid is only 300 km across, the data will be adequate to reveal its size, shape and albedo. If it is bigger, so much the better. The Voyager cameras will also monitor Nereid's light variation. The combined data set should give a better determination of its spin period and a clue to the orientation of the spin axis.

Spectral coverage of Nereid from Voyager will be limited, but sufficient to test whether the satellite has the peculiar colour characteristics suggested by the data reported in this issue⁶. Future JHK and higher-resolution infrared spectral measurements will be needed to begin sorting out Nereid's surface composition. Because the Voyager encounter with Nereid will be a distant one, no mass determination can be made. Nereid's mean density and internal make-up will remain unknown, as will, almost certainly, the satellite's origin. □

1. Cruikshank, D. P. & Brown, R. H. in *Satellites* (eds Burns, J. & Matthews, M. S.) 836–873 (Univ. Arizona, 1986).
2. Veillet, C. *Astr. Astrophys.* **112**, 277–280 (1982).
3. Harrington, R. S. & Van Flandern, T. C. *Icarus* **39**, 131–136 (1979).
4. Hartmann, W. K. *et al.* *Icarus* **47**, 333–341 (1981).
5. Peale, S. J. in *Planetary Satellites* (ed. Burns, J.) 87–112 (University of Arizona Press, 1977).
6. Schaefer, M. W. & Schaefer, B. E. *Nature* **333**, 436–438 (1988).
7. Hartmann, W. K., Cruikshank, D. P. & Degewij, J. *Icarus* **52**, 377–408 (1982).
8. Cruikshank, D. P., Hartmann, W. K. & Tholen, D. J. *Nature* **315**, 122–124 (1985).
9. Thomas, P., Veverka, J. & Dermott, S. in *Satellites* (eds Burns, J. & Matthews, M. S.) 802–835 (University of Arizona Press, 1986).

J. Veverka is in the Center for Radiophysics and Space Research, Cornell University, Ithaca, New York 14853, USA.

Fluid dynamics

Stirred but not mixed

S. Wiggins

In many circumstances the bulk flow of a fluid is more important for mixing and transport than the diffusive component. As a commonplace example, consider the mixing of cream with coffee. Experience shows that, to disperse the cream quickly and efficiently, it is much more effective to stir the mixture rather than to allow the cream to sit in the coffee and to rely on molecular diffusion. On page 419 of this issue¹, however, Ottino *et al.* present counterintuitive examples of stirring resulting in structured flow that isolates islands of unmixed fluid. Similar structures are observed in plots of the dynamical behaviour of certain nonlinear dynamical systems in which chaotic behaviour (corresponding to interaction of filaments) and quasi-periodic behaviour (corresponding to islands) can occur.

To understand how mixing occurs it is necessary to know how fluid particles move under the action of the fluid velocity field. It has become clear that extremely complicated fluid particle motions can occur in simple fluid dynamical circumstances such as stirring. For example, the motion of three point vortices in unbounded flow is integrable (well-behaved), but the motion of fluid particles near the vortices can be chaotic². Another example is Arnold–Beltrami–Childress flow³, a steady, spatially periodic solution of the Euler equations of inviscid-fluid mechanics (thus dynamically simple) for which the individual particle motions can

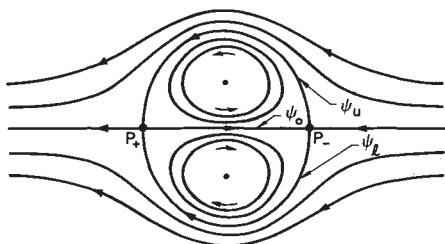
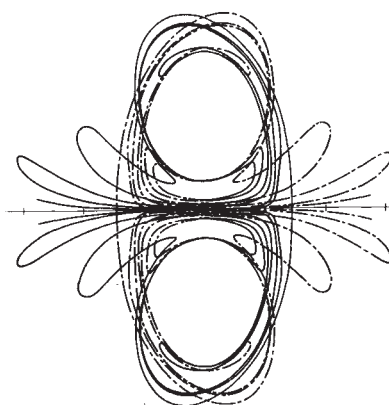


Fig. 1 The streamlines of the time-independent, two-dimensional, incompressible, inviscid flow produced by two counter-rotating point vortices of equal strength. P_+ and P_- are stagnation points (hyperbolic fixed points) for this flow (points of zero fluid velocity). Because fluid particles cannot cross streamlines, ψ_u and ψ_l bound a region of fluid which is forever trapped near the vortices making an island. In the context of dynamical-systems theory, ψ_u and ψ_l are examples of invariant manifolds: all initial conditions starting on the curve must always remain on the curve under the evolution of the dynamical system. In this case, ψ_u and ψ_l make the stable manifold of P_+ and the unstable manifold of P_- : fluid particles starting on ψ_u and ψ_l approach P_+ asymptotically for future times and P_- asymptotically for past times. The converse applies to ψ_u .

be chaotic. The flood of results in the theory of nonlinear dynamical systems, brought about by vigorous interest in the topic, offers the means to elucidate these problems, and provides a proper framework for the study of mixing and transport.

How are the dynamics of fluid particles and nonlinear dynamics related? If $\mathbf{v} =$

Fig. 2 A time-periodic strain-rate field is added to the flow in Fig. 1, making the velocity field time-dependent. The topology of the flow can appear deceptively complicated because different fluid particle trajectories may pass through the same point (although at different times). This difficulty is eliminated by studying the Poincaré map of the flow obtained by stroboscopically viewing fluid particle trajectories at time intervals equal to the period of the strain-rate field. The stagnation points P_+ and P_- of Fig. 1 persist as fixed points of the Poincaré map. The stable manifolds (solid lines) of P_+ and the unstable manifolds (broken lines) of P_- persist; but in the time-dependent flow, rather than coinciding, they can intersect in a countable infinity of discrete points forming a heteroclinic tangle. The breakup of the manifolds gives a mechanism for fluid trapped near the vortices to mix with the outer fluid. The intersection of the manifolds also provides the mechanism for chaotic fluid particle trajectories as shown in Fig. 3.



$(v_x(x,y,t), v_y(x,y,t))$ denotes the velocity field of a two-dimensional, incompressible flow, then the components of this field can be represented as $v_x = \partial\psi/\partial y$ and $v_y = -\partial\psi/\partial x$, where $\psi(x,y,t)$ is the 'stream-function'. The differential equations describing the motion of fluid particles (or equivalently, of passive tracer particles) in this velocity field are given by $dx/dt = \partial\psi/\partial y$, $dy/dt = -\partial\psi/\partial x$. To the nonlinear dynamicist, these are Hamilton's equations, where the streamfunction $\psi(x,y,t)$ plays the role of the hamiltonian function (Figs 1 and 2).

If the explicit time dependence of ψ is periodic, then the study of the continuous evolution described by the differential equations can be reduced to the study of a discrete time system, the 'Poincaré map' (Fig. 3) originally devised for periodically driven dynamical systems: a snapshot is taken each period of the system, and the mathematics involves the 'mapping' of points from one picture to the next. In this way, the theory of area-preserving maps of the plane — in dynamics, those of systems that are non-dissipative or energy conserving — can be brought to bear on problems of mixing and transport in fluid flows.

This theory provides the language to describe the important features of fluid flow. Kolmogorov–Arnold–Moser (KAM) theory⁴ and Aubry–Mather theory⁵, which describe sufficient conditions for quasi-

periodic orbits in nonlinear hamiltonian systems, provide information on the existence of flow structures that inhibit fluid transport. The Smale–Birkhoff homoclinic theorem⁶, which defines the criteria for chaotic behaviour in dynamical systems, prescribes the conditions necessary for chaotic fluid particle motions and hence enhanced mixing. The Liapunov exponent⁷, which describes the rate at which two trajectories starting at nearly the same point diverge, can be used to quantify the way fluid interfaces stretch.

In this issue¹, Ottino *et al.* describe two experiments which vividly demonstrate

these ideas: the flow in a rectangular cavity driven by two walls sliding periodically, and the flow between two eccentric cylinders rotating periodically, each giving a two-dimensional, time-periodic flow. By placing blobs of passive tracer in the flows, they can observe cases in which filaments of structured flow isolate islands of mixed fluid from one another. These structures are real analogues of the mathematical notions described above.

An important point is that the dynamics of mixing and transport are dominated by the hyperbolic points in the flow (see Fig. 1), and the associated stable and unstable manifolds (curves), rather than by other dynamical features. In particular, the stable and unstable manifolds are invariant filaments which can both inhibit and enhance the motion of fluid particles. Where they cross (intersect transversely) they promote the formation of Smale horseshoes: a small rectangle of fluid quickly becomes stretched, thinned and folded so that points originally close become widely separated and others are drawn close together (Fig. 3). Thus the intersections cause mixing of the fluid just as they cause chaotic behaviour in dynamical systems. Where an unstable and a stable manifold just fail to cross, they can act to form invariant regions that inhibit transport and mixing, termed 'incomplete horseshoes' by Ottino *et al.*¹.

Although only two relatively simple

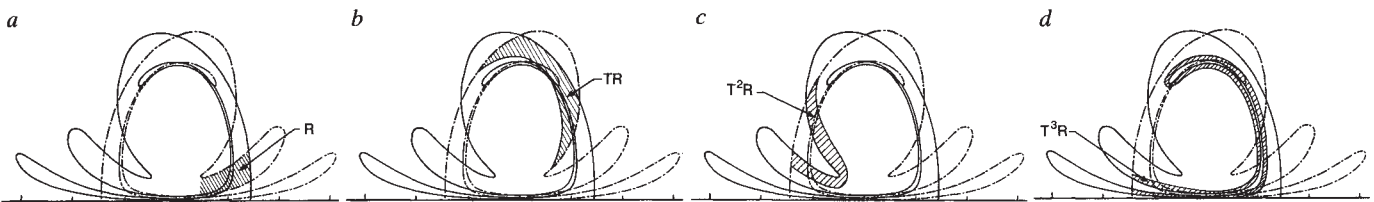


Fig. 3 Probably the best understood type of chaos arising in maps is the Smale horseshoe. A horseshoe arises when a region of the domain of the map is stretched, contracted and folded over itself in the shape of a horseshoe. The Smale-Birkhoff homoclinic theorem gives conditions on the existence of horseshoes based on the transverse intersections of the stable and unstable manifolds of fixed points of maps. *a-d*, A region labelled *R* from Fig. 2 evolves under the Poincaré map *T*. In *a* the horizontal boundaries of *R* are on pieces of the stable manifold of P_+ and a vertical boundary is on the unstable manifold of P_- . Because the manifolds are invariant, these parts of the boundary of *R* must remain on the manifolds. Repetitions of the operation *T*, *TR* (*b*), T^2R (*c*) and T^3R (*d*) stretch and fold *R*. T^3R appears to pass through *R* and is an example of a Smale horseshoe map.

flows are discussed in the paper, the important point is made that the mathematical structures (such as KAM tori, stable and unstable manifolds, Smale horseshoes) are generic in the sense that one expects them to occur in a large class of flows. This approach raises the possibility that much of the study of mixing and transport processes may be reduced to the study of certain building blocks.

The flows considered by Ottino *et al.* are two-dimensional and exhibit only a periodic dependence on time. Thus, although individual fluid particle motions can be chaotic, the overall flows are not turbulent in the classical sense. Nonlinear dynamics could have much to offer to studies of turbulent fluid flows, because there are large-scale organized motions in turbulent flows⁸. Unfortunately, theoretic-

cal understanding of these organized structures has lagged far behind the experiments. But nonlinear dynamics is concerned with the global topological structure of vector fields, an approach similar to the experimental framework which originally led to the discovery of the organized structures, and so offers a means to developing new theories.

The techniques of nonlinear dynamics are not yet sufficiently developed to answer all questions concerning mixing and transport in fluid flows. But, conversely, many questions in fluid dynamics point to areas where new mathematical concepts and techniques in nonlinear dynamics must be developed. For example, incompressible fluid flow must be reducible to a two-dimensional Poincaré map if the theory of area-preserving maps of the plane is to be applicable. But there is no recipe for this reduction for flows with non-periodic time-dependence. Therefore, similar techniques and structure theorems must be developed for flows with a more general time dependence which are ubiquitous in practice. Also, in technical applications it is often desirable to effect mixing and transport as quickly as possible with minimum effort and therefore new means to quantify rate processes are needed. Finally, the discussions of fluid particle motions from the point of view of nonlinear dynamics thus far have been for incompressible flows. In this case fluid volumes are preserved in the flow, thus excluding the possibility of attractors, sets of favoured final states. Compressible flows raise the possibility of 'strange attractor' fluid structures. Perhaps these real strange attractors could point the way to fruitful studies of their mathematical counterparts. □

Human AIDS virus not from monkeys

MANY African green monkeys show serological evidence of infection with a virus very similar to the human immunodeficiency virus (HIV), the AIDS virus. Several groups have reported that between 30 and 50 per cent of monkeys caught in the wild are seropositive for this virus, but these infected animals show no symptoms. Hayami's group in Tokyo¹ recently isolated from wild-caught monkeys the first 20 strains of the virus, called simian immunodeficiency virus (SIV_{agm}). The same group reports on page 457 of this issue² the complete nucleotide sequence of the first of these isolates. This sequence is no more similar to HIV-1 or HIV-2, the two human AIDS viruses, than each of the human viruses is to the other. In fact, Hayami's group will soon publish data showing that the SIV they isolated from the mandril (SIV_{mdr}) is equally closely related to the other three viruses, HIV-1, HIV-2 and SIV_{agm} (see the paper in this issue³).

The genomic structure of SIV_{agm} is very like the three other completely sequenced primate immunodeficiency viruses, HIV-1, HIV-2 and SIV_{mac} (refs 3, 4 and 5, respectively) even though the nucleotide sequences are not very similar. The genome contains open reading frames analogous to the seven main genes common to these viruses. SIV_{agm} shares with HIV-2 and SIV_{mac} the 'X' gene which is absent in HIV-1, but it lacks the 'R' gene that the other three HIV/SIVs share. The functions of X and R are unknown. The SIV_{agm} envelope protein, with its larger number of cysteine residues, could have a more complex structure than those of the other three viruses.

All these immunodeficiency viruses belong to the lentivirus subgroup of retroviruses. The only other simian lentivirus with a known nucleotide sequence⁴, SIV_{mac}, is

remarkably similar to HIV-2 and was originally isolated from Rhesus monkeys with an AIDS-like syndrome⁶. But the origin of SIV_{mac} and the closely related SIV_{mne} (ref. 7) is a mystery. It has been found only in colony-born Asian macaques in two primate centres in the United States; nobody has found any wild-caught Asian monkeys with serological evidence of SIV infection.

Contrary to SIV_{mac} and SIV_{mne}, the SIV_{agm} is not known to cause any disease in its natural host or in any other monkey, although many African green monkeys have serological evidence of SIV infection. It is obviously important to find out what makes SIV_{agm} non-pathogenic or, conversely, what makes the African green monkey resistant to SIV_{agm}. The fact that the SIV_{agm} sequence is so remarkably different from the human AIDS viruses indicates that the human viruses cannot have originated from African green monkeys in recent times, as had been predicted by many people. This claim was usually based on what was very probably a case of mistaken identity (refs 8, 9 and my News and Views article¹⁰). It cannot be excluded that these lentiviruses were present in the common ancestor of Old World primates and humans and have evolved along with them². Carel Mulder

- Ohta, Y. *et al.* *Int. J. Cancer* 41, 115-122 (1988).
- Fukasawa, M. *et al.* *Nature* 333, 457-461 (1988).
- Ratner, L. *et al.* *Nature* 313, 619-621 (1988).
- Guyader, M. *et al.* *Nature* 326, 662-669 (1987).
- Chakrabarti, L. *et al.* *Nature* 328, 543-547 (1987).
- Daniel M.D. *et al.* *Science* 228, 1201-1204 (1985).
- Benveniste, R.E. *et al.* *J. Virol.* 62, 2090-2101 (1988).
- Kestler, III, H.W. *et al.* *Nature* 331, 619-621 (1988).
- Essex, M. & Kanki, P. *Nature* 331, 621-622 (1988).
- Mulder, C. *Nature* 331, 562-563 (1988).

Carel Mulder is Professor of Pharmacology, Molecular Genetics and Microbiology at the University of Massachusetts Medical School, Worcester, Massachusetts 01655, USA.

- Ottino, J.M., Leong, C.W., Rising, H. & Swanson, P.D. *Nature* 333, 419-425 (1988).
- Aref, H. & Pomphrey, N. *Proc. R. Soc. A* 380, 359 (1982).
- Dombre, T. *et al.* *J. Fluid Mech.* 167, 353 (1986).
- Lichtenberg, A.J. & Leiberman, M.A. *Regular and Stochastic Motion* (Springer, New York, 1983).
- Mackay, R.S., Meiss, J.D. & Percival, I.C. *Physica* 13D, 55 (1984).
- Wiggins, S. *Global Bifurcations and Chaos-Analytical Methods* (Springer, New York, 1988).
- Oseledec, V.I. *Trans. Mosc. math. Soc.* 19, 197 (1968).
- Jimenez, J. (ed.) *The Role of Coherent Structures in Modeling Turbulence and Mixing* (Springer, New York, 1981).

S. Wiggins is in the Division of Engineering and Applied Science, California Institute of Technology, Pasadena, California 91125, USA.

Igneous geology

Causes of magmatic diversity

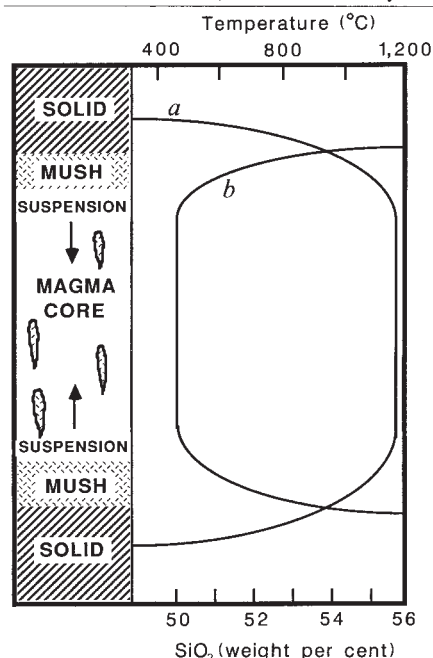
Bruce Marsh

WHAT Charles Darwin did for organisms he also did for volcanic rocks: he explained diversity using a simple evolutionary principle and gave magmatists a source of endless controversy. Black basalts, separated by some 25 weight per cent in SiO_2 content from pink granites, can progressively change, by continual separation of early forming crystals, into a much more refined granitic composition. This is fractional crystallization, the technique used by Curie to purify uranium ores. It is easily invoked in magmas because, without exception, magmatic crystals differ markedly in composition from their multicomponent parent melt. Solid-liquid density contrasts encourage relative motion and separation of phases to yield a liquid line of descent linking basalts to granites. This classical model has suffered greatly over the past twenty years, mostly from the findings of the new field of magma-physics studies, which suggest that magmas are highly turbulent and not conducive to crystal settling. The findings of Cox and Mitchell, reported on page 447 of this issue¹, indicate that the tide is turning back towards classical explanations. They show that sequential eruptions of crystal-laden and crystal-free basalts from the Deccan plateau in India are compositionally connected through their crystal content and density contrast.

Magmatic diversity has always been recognized as a complex problem. Huge Hawaiian magmatic systems, highly fluid and laced with sinkable crystals, produce only basalt. Tiny, sluggish island arc volcanic centres, barely able to sustain more than a few successive eruptions, show good diversity. Why? Part of the answer, until recently, seemed to be related to the convective dynamics of the magma itself. Large vats or magma chambers beneath volcanoes convect as they cool, crystallize and fractionate. The vigour of convection in these often sheet-like bodies depends on the rate of heat transfer, magma viscosity and the size of the system. Large, highly fluid, near-surface chambers would convect turbulently prohibiting crystal settling, whereas small, viscous, crystal-rich chambers would be largely stagnant, allowing efficient crystal fractionation.

Thickly ponded Hawaiian magma, large enough to constitute respectable magma chambers, drilled during 30 years of cooling, gives clear evidence that these bodies are not turbulently convecting, and may in fact be closer to stagnancy (see figure). These lava lakes cool as if domina-

ted by conduction², and small dribbles of fractionated liquid from the floor traverse the chamber without destruction by mixing³. The body crystallizes at the chamber roof and floor, by forming crusts that trap nearly all the remaining fractionated fluid. This cool, dense fluid would normally drive the magma convection were its viscosity insensitive to temperature. The hot inner parts of the body are



Schematic of an Hawaiian lava lake, 100 metres deep, some 10 years after emplacement¹. Crusts grow inwards from the top and bottom margins. A low crystallinity suspension zone and a higher crystallinity mush zone form the inner reaches of each crust. The innermost temperature (a) is still that of the initial body. The silica content (b) of the interstitial liquid, a measure of degree of fractionation, rises sharply within the suspension, mush and crust regions, but only inwards of the suspension zone is such liquid eruptable. Small dribbles of this refined melt leak up from the lower crust and reach the upper without convective mixing.

protected from the cooler country rock by these insulating blankets or crusts. Although rising gas bubbles and the downward growing crust can capture suspended crystals, most crystals initially in the rising magma settle out, despite convection.

New crystals grow near and within the propagating crusts, and are excluded from the hot inner core of the body where they remelt⁴. But these bodies do not produce large volumes of fractionated magma because the most fractionated liquids are trapped in the crusts, apart from the occasional small dribble that escapes from

the floor. The crusts growing on vertical side walls similarly capture viscous, refined melt as it flows sluggishly up through the porous crust to collect at the roof. So it is not convection, but crust, that inhibits fractionation.

Cox and Mitchell have looked at basalts in the Deccan plateau in Northern India which was magmatically active 65 million years ago. Thus rather than looking at existing magma chambers, they have to infer the magma dynamics from evolved basalts. By assuming that successive lavas are accurate samples or snapshots of the evolution of an associated magma chamber, they find that the initial crystal content of the magma was so high that once the crystals had settled out, the magma produced a highly fractionated liquid. Further fractionation of these crystal-free liquids would have been impossible because all later crystals grow near crust which traps them and prohibits crystal settling.

The nature of the convective state is intimately tied to the dynamics of solidification⁵. Convection in magma chambers does not depend on the full thickness of the system, as it does in layered systems heated from below⁶ which leads to turbulence, but depends on the ratio of the thermal diffusivity to the solidification rate. Rapid, early growth of crust inhibits convection, which sets in later, is gentle and hugs the upper crust, not reaching the floor at all. Cooling at the base further encourages stagnancy near the floor⁷; there is no basal inward heat flux to stimulate convection⁸. The initial inner crystals are free to settle or remain in suspension. Even if the body is suddenly heated from below by influx of new, denser magma and convection becomes vigorous, crystal settling continues⁹.

While models of fractionation have varied between extremes, most workers have ploughed along using the time-honoured edict 'when in doubt, settle it out'. As it turns out, this is only partly true in the sense that it is necessary to know how and where to 'settle it out'. The appreciation of the subtle balance between cooling, solidification, convection, phase equilibria, rheology and crystal sorting greatly increases our understanding of magma dynamics and magmatic diversity. □

1. Cox, K.G. & Mitchell, C. *Nature* **333**, 447-449 (1988).
2. Peck, D.L., Hamilton, M.S. & Shaw, H.R. *Am. J. Sci.* **277**, 415-437 (1977).
3. Helz, R.A. *Geochem. Soc. spec. publ.* No. 1, 241-258 (1987).
4. Mangan, M.T. & Marsh, B.D. *Am. geophys. Un. Trans.* **69**, 526 (1988).
5. Smith, M.K. *J. Fluid Mech.* **188**, 547-570 (1988).
6. Huppert, H.E. & Sparks, R.S.J. *Contr. Min. Petrol.* **75**, 279-289 (1980).
7. Jaupart, C., Brandeis, G. & Allegre, C.J. *Nature* **308**, 535-538 (1984).
8. Carrigan, C.R. *Geophys. Res. Lett.* **14**, 915-918 (1987).
9. Martin, D. & Nokes, R. *Nature* **332**, 534-536 (1988).

Bruce Marsh is in the Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, Maryland 21218, USA.

Neutron physics

Half life defies measurement

J. Byrne

THE β -decay of a free neutron into a proton, electron and neutrino¹ is an extremely rare event compared with other sub-nuclear reactions. The first measurement² of the neutron lifetime, a parameter of wide theoretical significance, gave the quite precise value $\tau = 1,013 \pm 26$ s. Subsequent experiments (Fig. 1) failed to sustain confidence in this result, revealing

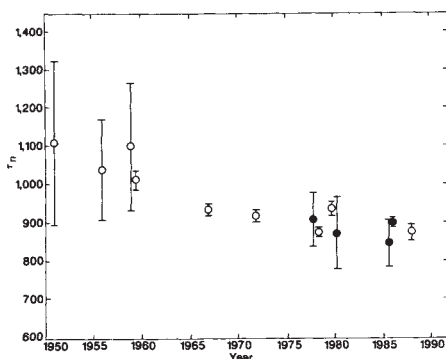


Fig. 1 The decrease in measured values of neutron lifetime τ_n . Open circles, estimates based on the flux of decay particles from neutron decays in neutron beams. Closed circles, estimates based on number of surviving neutrons in magnetic or material bottles.

discrepancies and uncertainties in the data which have been scrutinized theoretically³ and experimentally⁴. Thus a new redetermination of τ_n reported by J. Last *et al.* from the Institut Laue-Langevin at Grenoble⁵, which proposes the value $\tau_n = 876 \pm 21$ s, arouses great interest, not only because it seems to confirm the consistent decline over many years in the measured values of τ_n , but also because it lays claim to no significant improvement in accuracy compared with its immediate predecessors. Although it can hardly be contemplated that the longevity of neutrons is less today than in former times, it is evident that τ_n has been seriously overestimated in the past, and it is interesting to enquire why this is so, and whether the latest measurement is any more reliable.

The neutron lifetime is of crucial importance for theories of the nuclear weak interaction, astrophysics and cosmology. In the context of weak interactions, precise measurements of the weak polar-vector (g_V) and axial-vector (g_A) coupling constants that govern nuclear β -decay can be made only by the study of free-neutron β -decay, which is independent of nuclear structure. Furthermore, because an independent value of g_V can be derived from the measured rate of superallowed pure Fermi β -transitions in which nuclear-structure effects are minimal and calculable, the possibility exists for a fundamental cross-checking of current theory⁶. The basic proton-proton cycle of energy generation in the Sun proceeds at a rate which is proportional to g_A^2 with obvious implications for the solar-neutrino problem⁷. In cosmology, τ_n is closely related to the rate at which helium is formed in the early Universe, and is important in the determination of the number of distinct types of light neutrino⁸.

To illustrate the imprecision of our knowledge of the neutron lifetime, we may compare the existing accuracy of about 2.5 per cent with the corresponding figures of 1.8×10^{-3} , 0.09 and 0.8 per cent for the muon, pion and the lambda particles respectively. Like the lambda, the neutron is one of seven baryons closely related to the proton whose decay parameters are all connected by Cabibbo's theory of weak interactions⁹, but whose lifetimes, excluding τ_n , are about 10^{-10} s. The fact that the neutron lives 10^{13} times longer than any of its partners (except the stable proton) is solely a reflection of the near identity of the neutron and proton masses.

The truth is that β -decay is a very rare event in the career of a neutron, and it is precisely because of this that τ_n is extraordinarily difficult to measure with any accuracy. No more than one neutron will decay per second per centimetre cubed in the neutron flux from a typical reactor,

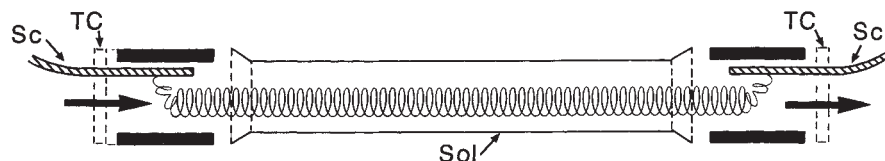


Fig. 2 Schematic of the experiment of J. Last *et al.*⁵. A pulsed beam of cold neutrons (arrows) passes through the superconducting solenoid (Sol). Electrons emitted in β -decays follow helical paths around the magnetic field lines to the scintillators (Sc) where they are detected. The shaped field tapers towards the scintillators to prevent the magnetic-mirror effect which would trap electrons. Timing information can identify events in which an electron is back-scattered from one scintillator to be detected a second time by the other, and the event is counted just once at the first detection. TC, magnetic trim coil.

100 years ago

THE following incident is from the trial of the great patent case, Edison and Swan Electric Light Company *v.* Holland and others, now proceeding in the High Court of Justice. On May 16, Prof. James Dewar was under examination. A small crucible was produced and handed to the witness, who said: In the crucible I have, with Mr. Gimingham, carbonized filaments in the precincts of the court, using no packing and no luting of any description. Sir Horace Davey urged that this did not arise out of the cross-examination. Mr. Justice Kay said it should have been produced in the examination-in-chief.

Mr. Justice Kay: I am very much disgusted. I am here trying all I can to understand the case, and this is clearly an attempt to mislead. I am greatly disgusted.

Prof. Dewar: I have no desire to mislead your lordship. I have stated that this was a mere experiment. I did not produce it. It was put to me.

Mr. Justice Kay: You may stand down.
From *Nature* 38, 114; 31 May 1888.

and unless elaborate precautions are taken, the demise of this one unlucky neutron is very difficult to detect against the intense radiation background generated when the surviving neutrons are captured locally in strong nuclear reactions. As a result it is very easy to lose sight of genuine neutron decays in the confusion of spurious background events, with the consequence that τ_n is overestimated.

Two quite different methods have been used in the past to determine τ_n (ref. 4). The first makes use of ultracold neutrons of energy less than 10^{-7} eV stored in magnetic or material bottles, and τ_n is determined by recording the number $N_n(t)$ of neutrons surviving to time t , and applying the exponential law of radioactive decay $N_n(t) = N_n(0) \exp(-t/\tau_n)$. The second method, which is used in the new measurement by Last *et al.*⁵, relies on the differential form of the decay law $dN_n(t)/dt = -N_n(t)/\tau_n$. In the new experiment, $-dN_n/dt$ is determined by counting electrons of energy in the range 0–0.78 MeV created by neutron decays within a 2.1-m length of beam containing N_n neutrons. These electrons are confined to move in helical paths of less than 0.5 cm orbit radius by a 1.6-tesla longitudinal magnetic field, so shaped that electrons are channelled into one of a pair of plastic scintillators located at each end of, but outside, the neutron beam (Fig. 2). This system provides a counter for all the decay events that occur within the specified active volume.

The use of a pulsed rather than a continuous neutron beam is a novel feature of this experiment. This technique eliminates the spread of neutron velocities from the equation, because all the neutrons in the pulse are contained within a 1.5-m length of beam and the number N_n of neutrons in this active volume is determined by absorbing the whole pulse in thick ^{59}Co or ^{197}Au targets. The resultant

activity is then measured to high accuracy in five independent laboratories.

This procedure avoids many of the problems which arise in the conventional method for determining N_n which requires that the active volume and the space-averaged neutron density within that volume be assessed separately. The use of electron rather than proton counting is, however, a severe weakness of the technique, because only 34.85 per cent of the integrated electron spectrum is resolved unambiguously from the detector noise, and the remaining fraction has to be deduced by a complex calibration procedure. Thus, even though the new value is almost identical with a result published 10

years ago¹⁰, which was claimed to be twice as accurate as the present one, it is difficult to avoid the conclusion that the neutron lifetime problem has still to be solved. □

1. Snell, A.H. & Miller, L.C. *Phys. Rev.* **74**, 1217–1218 (1948).
2. Sosnovski, A.N. *et al.* *Nucl. Phys.* **10**, 395–404 (1959).
3. Wilkinson, D.H. *Nucl. Phys.* **A377**, 474–507 (1982).
4. Robson, J.M. *Contemp. Phys.* **24**, 129–141 (1983).
5. Last, J. *et al.* *Phys. Rev. Lett.* **60**, 995–998 (1988).
6. Carney, A.S., Deutsch, J. & Holstein, B.R. *Phys. Rev. D* (in the press).
7. Bahcall, J.H. *et al.* *Rev. mod. Phys.* **54**, 767–799 (1982).
8. Tayler, R.J. *Rep. Prog. Phys.* **43**, 253–299 (1980).
9. Cabibbo, N. *Phys. Rev. Lett.* **10**, 531–533 (1963).
10. Bondarenko, L.N. *et al.* *JETP Lett.* **28**, 303–307 (1978).

J. Byrne is in the School of Mathematical and Physical Sciences, Sussex University, Brighton BN1 6WF, UK.

Microbiology

Regulation of sporulation

Jeffery Errington

BACILLUS subtilis, for many years intensively studied as a model system for development and differentiation, but having recently fallen on hard times, may be about to make a comeback. This was the take-home message of a recent meeting*, at which there was a general air of optimism that the mechanisms of spatial and temporal control of sporulation gene expression will soon be understood. Almost all the genes controlling development (*spo*) have now been cloned (see ref. 1, for example, and my own work) and the sequences of many have been determined.

B. subtilis, which grows and divides much like any other rod-shaped bacterium in relatively rich media, undergoes a 7-hour-long, two-cell differentiation process when starved. The immature prespore cell develops via a series of intermediate morphological stages into a highly resistant, dormant spore within the cytoplasm of a larger 'mother cell' (see Fig. 1). The two cells, both of which contain single intact chromosomes, are derived by the asymmetrical division of a single undifferentiated parental cell. What is the nature of the mechanism that determines differential gene expression? Distinct pathways of gene activation have now been established in each of the cellular compartments (see Fig. 2), mainly by the study of epistasis relationships using *spo-lacZ* fusions¹. But early in the process, at least some of the important regulatory genes, for example the gene encoding σE (formerly called σ_{29}), seem to be expressed in both differentiating cells (W.G. Haldenwang, University of Texas Health Science Center, San Antonio). Clearly, somewhere in between is a mechanism for ensuring certain key regulatory genes are activated in only one or

other of the compartments. Several good candidates for the determinants of spatial expression have now been identified.

One of the new DNA sequences reported at the meeting is particularly notable. The product of the *spoIIIM* gene turns out to be very similar to an important cell-division gene, *ftsA*, from *Escherichia coli* (T. Leighton, University of California,

becoming increasingly clear that there are distinct differences from the paradigms of *E. coli*. For example, *B. subtilis* RNA polymerase is now known to display a remarkable heterogeneity. Two new forms of sigma factor (σ , the subunit of RNA polymerase that determines promoter specificity), both involved in sporulation, were reported, bringing the total number of known forms in this organism to nine (ref. 4). The sigma factor σG is the product of a newly identified gene, *spoIIIG* (Y. Kobayashi, Hiroshima University; P. Setlow, University of Connecticut Health Center, Farmington; P. Stragier, Université Paris-Sud, Orsay), and seems to be involved in spore-specific gene expression. Two other loci implicated in this mechanism, *spoIIIA* and *spoIIIE*, have also been identified and are being characterized. The *spoIIIA* locus is involved in the regulation of *spoIIIG* expression; *spoIIIE* expression is independent of *spoIIIA* and its product appears to act in parallel with σG . The *spoIIIE* gene is particularly interesting because, although it is required specifically for the expression of genes in the spore compartment, it turns out to be expressed during vegetative growth, before the formation of the spore septum that separates the differentiating cells (our unpublished results; see Fig. 2). Presumably, its function is mediated by a

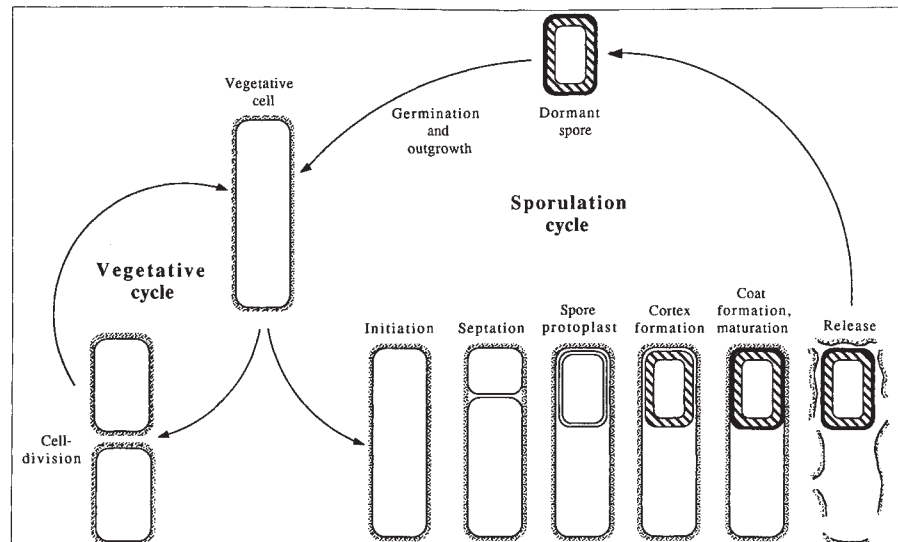


Fig. 1 Simplified life cycle of *Bacillus subtilis*, illustrating the major stages in development.

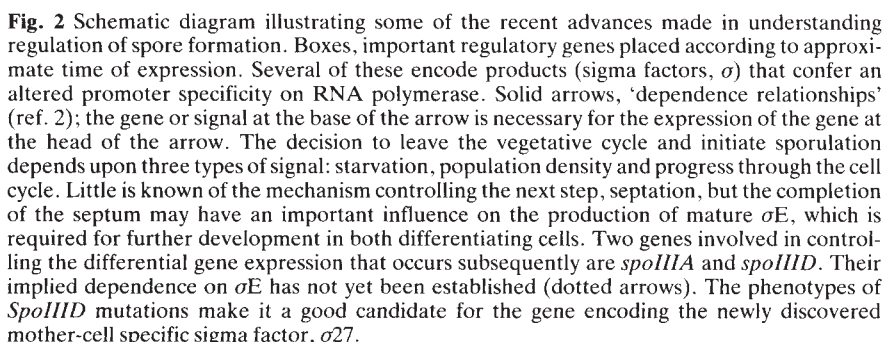
Berkeley). This is an important finding because it reinforces the old belief that the control of sporulation is intimately linked to the cell-division cycle³. It could also reawaken interest in the study of *B. subtilis* sporulation as a model for cellular septation and division. Unlike normal division events, the sporulation septum can be completely blocked in mutants without loss of viability.

But *Bacillus* is not just interesting as a model for cell division and differentiation. As the mechanisms of gene regulation begin to unfold in this organism, it is

partitioning or differential activation mechanism linked to septation events.

Another new sigma factor, of relative molecular mass 27,000 (27K), was reported by L. Kroos (R. Losick's laboratory, Harvard University). What is remarkable about this new protein is that its promoter specificity (the type of promoter sequence from which it directs transcription) can be completely changed *in vitro* by the action of a 14K protein (see Fig. 2). Kroos and Losick speculate that this change in promoter specificity is coupled to morphogenic events that occur late in

* The 10th International Spores Meeting Woods Hole, Massachusetts, 23–27 March 1988.



Finally, much work is being done on the genes that control the initiation of sporulation — the decision to divide vegetatively or enter the sporulation cycle. In this area it is clear that at least one specialized form

1. Errington, J. & Jones, D. J. *gen. Microbiol.* **133**, 483-492 (1987).
2. Mandelstam, J. & Errington, J. *Microbiol. Sci.* **4**, 238-244 (1987).
3. Dawes, I.W. *et al. Nature* **230**, 567-569 (1971).
4. Helmann, J.D. & Chamberlin, M.J. *A. Rev. Biochem.* (in the press).
5. Stragier, P. *et al. Cell* **52**, 697-704 (1987).
6. Dubnau, E. *et al. J. Bact.* **170**, 1054-1062 (1987).
7. Trach, K.A. *et al. Proc. natn. Acad. Sci. U.S.A.* **82**, 7260-7264 (1985).
8. Ninfa, A.J. & Magasanik, B. *Proc. natn. Acad. Sci. U.S.A.* **83**, 5009-5013 (1986).

David Jones

Effects of microwaves on cells and molecules

SIR—For a paper stating that “the literature on biological effects of low-level microwave energy remains shot through with ambiguity”, Foster and Pickard’s commentary on the risks of microwaves¹ is wide of the mark and unreasonably denigrates the efforts of the many who have established structural and functional substrates for essential aspects of these interactions at cellular and subcellular levels². They chose as examples of current knowledge three unrelated observations that are not now, nor were ever, in the mainstream of studies of mechanisms of low-level radiofrequency (r.f.) and microwave biological effects: the auditory effect; rate changes in the isolated frog heart; and altered permeability in the blood–brain barrier.

Foster and Pickard ignore a wealth of findings that relate to the role of cell membranes in transductive coupling of electrical and chemical signals from the outside to the inside of the cell³; in interactions at cell surfaces between weak electromagnetic fields (extremely low-frequency (ELF) fields and r.f./microwave fields amplitude- or pulse-modulated at ELF frequencies) and the gamut of natural and artificial chemical molecules that stimulate cell surfaces at specific receptor sites; in the occurrence of these interactions only in certain bands of modulation frequencies and at certain field intensities, constituting ‘windowed’ interactions⁴; and in the clear evidence that these processes involve enormous amplification between initial energies of the first binding triggers and their ultimate actions within cells⁵. Beyond the evidence for these modulation-dependent effects, further research is needed to evaluate the nature and extent of possible resonant interactions between biomolecules and millimetre microwave fields.

Evidence for cell membrane interactions with modulated r.f./microwave fields is firmly based on modern cell and molecular biology. Of prime importance is the demonstration that these fields sensitively modify calcium binding at cell

surfaces and modulate a host of calcium-dependent intracellular enzyme mechanisms⁶. These intracellular markers of events initiated at cell surfaces regulate the flow of intracellular messages (protein kinases), cell metabolism and cell growth.

In less than a century, the frontier of our understanding has moved from tissues to cells to molecules. The use of imposed fields will now reveal the essential importance of biological organization at the atomic rather than the molecular level, and in physical rather than chemical terms; with coherent states between adjacent molecular electric charges and enormous cooperativity in energy release by very weak triggers as the physical essence of living matter^{7,8}.

Studies of chemical tumour promoters are providing new knowledge about their disruption of normal communication between cells with onset of unregulated growth⁹; bioelectromagnetic research is producing clear evidence of joint cell-surface actions of tumour promoters with ELF and modulated r.f. fields¹⁰. The union of these two disciplines has initiated the first significant new approach to tumour formation in 75 years — an approach that directs attention to dysfunctions in inward and outward streams of signals at cell membranes, rather than to damaged DNA in cell nuclei¹¹.

ROSS ADEY

*Jerry L. Pettis Memorial Veterans’
Hospital,
11201 Benton Street,
Loma Linda, California 92357, USA*

Training the brain using neural-network models

SIR—Recent discussions in *Nature*^{1,2} and elsewhere³ highlight the potential of parallel distributed processing (PDP) for modelling various aspects of cognition. The PDP approach, however, is circumscribed by the need of a network to be informed of its output error so as to modify hidden units through back-propagation. Without this performance, feedback training of the network cannot occur. Consequentially, there exists a paradox within PDP; for a network to learn anything it must in some way already possess a means of knowing what is and is not a correct output — but the ability to do this is precisely what the network has been created to explain.

This paradox does not affect present-day running of PDP models as realized on computers. PDP experimenters train them with thousands of example inputs with correction feedback. But connectionist models are claimed to be able to explain actual brain processes. This raises a question: what takes the place of the training provided by PDP experimenters for the brain’s neural networks?

One possibility is that the brain tutors its own neural networks through a bootstrapping process. Cognition could involve development of easily acquired tutoring processes of limited efficiency and effectiveness which in turn tutor later and cognitively more effective and efficient processes.

The development of reading may provide such an example. Children pass through several stages in learning to read English. Initially — often before entering school — children recognize some frequently occurring words using visual ‘logographic’ characteristics^{4,5}. Most children then quickly learn to use the phonological information present in written words⁶. Later, the use of phonological information loses importance as a more

abstract non-phonological letter sequence⁵ information takes its place. The limited vocabulary of logographically recognizable words could provide the examples needed to train the phonological neuronal network. This relationship could occur between the phonological and orthographic neuronal networks with the phonological decoding of written words functioning not only to identify them but also to tutor the development of hidden units in the orthographic network. This might explain why phonological decoding skills predict reading success⁷ even though mature reading in English is not itself phonological.

PDP has been called behaviourism in computer’s clothing⁷. While this analogy is misleading, it is true that the PDP account of learning requires a kind of conditional reinforcement in the form of adjustment of hidden units by back-propagation. The circumstances which could optimize the availability of information for this reinforcement — its ‘tutorability’ — is an aspect of PDP that so far has received little attention. But it could be important to our understanding of the role of PDP models in the development of human cognition.

JOHN R. SKOYLES

*Department of Psychology,
University College London,
Gower Street,
London WC1E 6BT, UK*

1. Foster, K.R. & Pickard, W.F. *Nature* **330**, 531–532 (1987).
2. *Microwave News* **8**, 2–3 (1988).
3. Fletcher, W.H., Byus, C.V. & Walsh, D.A. in *Regulation of Testicular and Ovarian Function* (ed. Mahesh, V.) 299–323 (Plenum, New York, 1987).
4. Blackman, C.F., Benane, S.G., House, D.E. & Joines, W.T. *Bioelectromagnetics* **6**, 1–11 (1985).
5. Adey, W.R. *Bioelectrochem. Bioenergetics* **15**, 447–456 (1986).
6. Adey, W.R. *Neurochem. Res.* **13**, 671–677 (1988).
7. Adey, W.R. & Lawrence, A.F. eds *Nonlinear Electrodynamics in Biological Systems* (Plenum, New York, 1988).
8. Frohlich, H. (ed.) *Biological Coherence and Response to External Stimuli* (Springer, Heidelberg, 1988).
9. Yamasaki, H. in *Nongenotoxic Mechanisms in Carcinogenesis* (eds Butterworth, B.E. & Slaga, T.J.) 297–307 (Cold Spring Harbor, Banbury Report 25, 1987).
10. Byus, C.V., Pieper, S. & Adey, W.R. *Carcinogenesis* **8**, 1385–1389 (1987).
11. Trosko, J.E. *Eur. J. Cancer clin. Oncol.* **23**, 599–601 (1987).

1. Anderson, A. *Nature* **331**, 657–658 (1988).
2. Rumelhart, D., Hinton, G.E. & Williams, R. *Nature* **323**, 533–536 (1986).
3. Rumelhart, D. et al. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Vols 1 and 2* (MIT, Cambridge, Massachusetts, 1986).
4. Seymour, P. & Elder, L. *Cogn. Neuropsychol.* **3**, 1–36 (1986).
5. Frith, U. in *Surface Dyslexia* (eds Patterson, K.E., Marshall, J.C. & Coltheart, M.) 301–330 (Erlbaum, London, 1986).
6. Bryant, P. & Bradey, L. *Children’s Reading Problems* (Blackwell, Oxford, 1985).
7. Papert, S. *Daedalus* **117**, 1–17 (1988).

Similarity of nucleotide sequences at splicing sites

SIR—We report an intriguing similarity between a sequence involved in processing of mitochondrial RNA and the consensus sequence at the 5'-splice site of introns in transcripts of nuclear genes.

In *Aspergillus nidulans*, synthesis of mature mitochondrial RNA occurs by cleavage and end-trimming of large precursor molecules¹⁻³. With one exception there are no significant similarities of sequence or structure between different processing sites. The exception concerns the sites where cleavages produce the 5'-termini of the mature 16S and 23S mitochondrial rRNA molecules. These sites lie in conserved 18-nucleotide motifs:



S1 nuclease analysis places the positions of the cleavage events in the quadruplet 5'-AGGU-3' indicated by the horizontal line.

The motif 5'-AGGU_A-3' also forms part of the consensus sequence for the 5'-splice junctions of introns in transcripts of nuclear genes⁴:



The mitochondrial processing sequence therefore includes six adjacent nucleotides that conform with the splice junction consensus. Furthermore, cleavage of the mitochondrial sequence may occur at the same relative position as cleavage of the splice site.

It is unlikely that the similarity has arisen purely by chance. Conceivably, components of the *A. nidulans* nuclear splicing machinery (snRNA, for instance) also have unsuspected roles in processing mitochondrial RNA precursors. This attractive hypothesis is tempered by the fact that the 5'-splice junction consensus for *A. nidulans* nuclear transcripts differs from the general version shown above in that the exonic 5'-CAG-3' sequence is absent⁵. Alternatively, the core sequence 5'-AGGU-3' may have an inherent disposition towards cleavage which aids RNA processing in both systems.

T.A. BROWN

Department of Biochemistry and Applied Molecular Biology,
UMIST,
Manchester M60 1QD, UK

N.J. DYSON

Amersham International,
Amersham HP7 9LL, UK

1. Dyson, N.J. thesis, Univ. Manchester (1988).
2. Brown, T.A. in *Gene Structure in Eukaryotic Microbes* (ed. Kinghorn, J.R.) 141-162 (IRL, Oxford, 1987).
3. Dyson, N.J. *et al.* *Gene* (submitted).
4. Watson, J.D. *et al.* *Molecular Biology of the Gene* 4th edn (Benjamin/Cummings, Menlo Park, 1987).
5. Gurr, S.J., Unkles, S.E. & Kinghorn, J.R. in *Gene Structure in Eukaryotic Microbes* (ed. Kinghorn, J.R.) 93-139 (IRL, Oxford, 1987).

Incubation period of AIDS in haemophiliacs

SIR—Ekert's assertion¹, which has led to discussion^{2,3}, that haemophiliacs infected with human immunodeficiency virus (HIV) may carry a tenfold-lower risk of developing AIDS than HIV-infected individuals from other risk groups, is at odds with our own findings. As we reported at the recent European Conference of Clinical Aspects of HIV Infection, we are caring for a cohort of 395 HIV-infected haemophiliacs, of whom 38 (10%) have developed AIDS so far. And in two large US cohorts⁴, 12 and 11% of HIV-infected haemophiliacs had developed AIDS by 1987. Thus it can be stated that the incidence of AIDS in HIV-infected haemophiliacs is not as low as proposed by Ekert.

With regard to the incubation period, information is difficult to obtain as the exact time of seroconversion is not known for most of the patients and as the number of HIV-infected patients in other risk groups is difficult to estimate³. During a mean follow-up period of 10 months, however, the disease progressed in 78 of 204 of our patients (38%) according to the CDC-classification, so one must assume that the incidence of AIDS in this cohort will increase further. This is in accordance with Goedert *et al.*⁵, who could not find a significant difference in the incidence of AIDS in cohorts of male homosexuals, intravenous drug-abusers and haemophiliacs.

T. KAMRADT
D. NIESE
B. KAMPS

Medizinische Universitätsklinik Bonn,
Klinische Immunologie,
S.-Freud-Strasse 25,
D-5300 Bonn 1, FRG

1. Ekert, H. *Nature* **329**, 494 (1987).
2. Turner, M.J., White, J.D. & Souter, W.P. *Nature* **330**, 702 (1987).
3. Rees, M. *Nature* **332**, 312 (1988).
4. The Pennsylvania AIDS Surveillance Study Group *Blood* **70**, 1208-1210 (1988).
5. Goedert, J.J. *et al.* *Science* **231**, 992-995 (1986).

Inheritance of insulin-dependent diabetes

SIR—Todd *et al.*¹ have presented data to suggest that inheritance of insulin-dependent diabetes mellitus (IDDM) through the HLA region is recessive, that variation at the DQ locus is more important than that at the DR locus in IDDM pathogenesis, and further that variation at a single amino acid residue (position 57 of the DQ_β chain) protects an individual from IDDM. We show here that recessively inherited predisposition to IDDM, due to a dominant protective effect from aspartate occupying residue 57, may be a misleading simplification.

Although the HLA-linked IDDM predisposition was initially described

as recessive, investigators have since proposed more complex and realistic models^{2,3} which take into account the following observations: greatly increased risk of IDDM in DR3/4 heterozygotes; a tendency to dominant behaviour of DR4 in the absence of DR3, and recessive behaviour of DR3 in the absence of DR4; increased risk associated with DR1 in many studies; and the pronounced protective effect of DR2 (ref. 4). A picture is gradually emerging of a scale of susceptibility on which every allele, indeed almost every genotypic combination, occupies a distinct place. Furthermore, it is now understood that DR3 and DR4 play differing roles in IDDM pathology, suggesting that IDDM is a heterogeneous disease⁵.

It is difficult to reconcile the hypothesis of Todd *et al.* of susceptible and protective haplotypes based on a single DQ_β residue with the following observations. First, the DR7 haplotype, like the susceptibility haplotypes DR1, DR3 and DR4, lacks aspartate at position 57 of the DQ_β molecule. The theory of Todd *et al.* thus predicts DR7 to be a high-risk haplotype, whereas in fact it confers no additional risk to IDDM. They suggest this discrepancy might be explained by the unusual structure of the DQ_α chain encoded in the DR7 haplotype, and yet this structure is very similar to the DQ_α chain encoded in the susceptible DR4 haplotype⁶. Second, in agreement with other studies, Todd *et al.* report that ~10% of IDDM patients carry the protective aspartate residue in DQ, so that susceptibility to IDDM due to the absence of this residue could not be strictly recessive. Third, the statements of Todd *et al.* that the DR locus is not involved in IDDM are contradicted by recent experiments on NOD mice, which lack the DR equivalents of class II MHC molecules (the I-E genes). As Todd *et al.* point out, NOD mice, which spontaneously develop diabetes, lack aspartate at position 57 of the I-A_β locus (equivalent to the human DQ_β), unlike other mouse strains. Nishimoto *et al.*⁷, however, were able to render the NOD mice non-diabetic by introducing the I-E genes, an experiment which suggests a role for the DR molecules in IDDM pathogenesis.

Current knowledge of the genetic contributions of the HLA region to IDDM strongly hints at several interactions: interaction within epitopes expressed on class II molecules, within single genes, within and between haplotypes, and finally interactions with unlinked loci and the non-genetic environ-

1. Todd, J.A. *et al.* *Nature* **329**, 599-604 (1987).
2. Rotter, J.I. *et al.* *Diabetes* **32**, 169-174 (1983).
3. Louis, E.J. & Thomson, G. *Diabetes* **35**, 958-963 (1986).
4. Thomson, G. *Am. J. hum. Genet.* **36**, 1309-1317 (1984).
5. Ludvigsson, L. *et al.* *Diabetologia* **29**, 207-210 (1986).
6. Moriuchi, J. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **82**, 3420-3424 (1985).
7. Nishimoto, H. *et al.* *Nature* **328**, 432-434 (1987).

ment. The awareness of one distinctive amino-acid residue might enrich our understanding of the panoply of HLA interactions in causing IDDM, but the facts offered, even from the laboratory of molecular biology, must fit into existing knowledge.

WILLIAM KLITZ

Department of Genetics,
University of California,
Berkeley,
California 94720, USA

Mantle and crustal Ce/Nd isotope systematics

SIR—Tanaka *et al.*¹ recently examined the degree of coupling of Ce and Nd isotope ratios in crustal and mantle systems. Based on theoretical considerations and a limited database, they argued that light rare-earth element (LREE)-enriched rocks (mostly continental plutons) display Ce and Nd isotope covariation along a "mantle array" (*sic*), whereas LREE-depleted rocks (mostly oceanic volcanics) have dispersed Ce and Nd isotope compositions which do not form an array. However, my own considerations of this question²⁻⁴ lead me to different conclusions.

The degree of correlation expected between mantle Ce and Nd isotope systematics may be deduced by comparison with Sr/Nd isotope behaviour. The Rb/Sr and Sm/Nd systems are very disparate in geochemical character, but oceanic volcanics define a clear Sr/Nd isotope mantle array for sources depleted relative to bulk earth. This is ascribed to somewhat coherent behaviour of incompatible elements during mantle melting events, and to convective mixing between mantle reservoirs⁵. Nevertheless, incompatible element ratios in erupted products are not an accurate guide to mantle source composition^{6,7}.

The greater chemical coherence within the LREE should lead to a substantially better Ce/Nd isotope mantle array than for Sr/Nd. Furthermore, the Mid-Ocean Ridge Basalt (MORB) source, which is the most LREE-depleted reservoir, is also the most well-mixed.

Ce/Nd isotope data for basic volcanics^{1,3} are plotted in Fig. 1a. The diagonal box defines a mantle array, delimited by average 2 sigma within-run precision in Ce isotope ratio on either side of the best fit regression of ocean island basalt (OIB) analyses³. Most of the data of Tanaka *et al.* are remarkably consistent with this array. Only three points, shown by hatched error bars, lie outside 2σ_m error limits of it.

Within-run precision is a good indicator of Nd isotope reproducibility, but may underestimate the uncertainty of some Ce isotope analyses due to a greater susceptibility of systematic errors¹. These may result from the large number of potential

mass interferences, susceptibility of the ¹⁴⁰Ce/¹⁶O peak tail to vacuum conditions, amplifier memory effects in dynamic peak switching analysis, and amplifier gain uncertainties in static analysis. Hence I shall examine the outliers from the mantle array to see if other isotope systems support their Ce isotope compositions, or whether these might be attributed to larger analytical errors.

Two points in Fig. 1a are from nearby localities on the Mid-Atlantic ridge. Both lie perfectly on the Sr/Nd isotope mantle array^{1,7,8}, but one of them has an unusual Ce isotope ratio. Similarly, a sample dredged from the East Pacific Rise lies off the Ce/Nd array, but comes, from a locality that yielded typical MORB Pb isotope ratios⁹. Finally, two samples from Iceland have similar Nd isotope compositions, but one has a Ce isotope ratio close

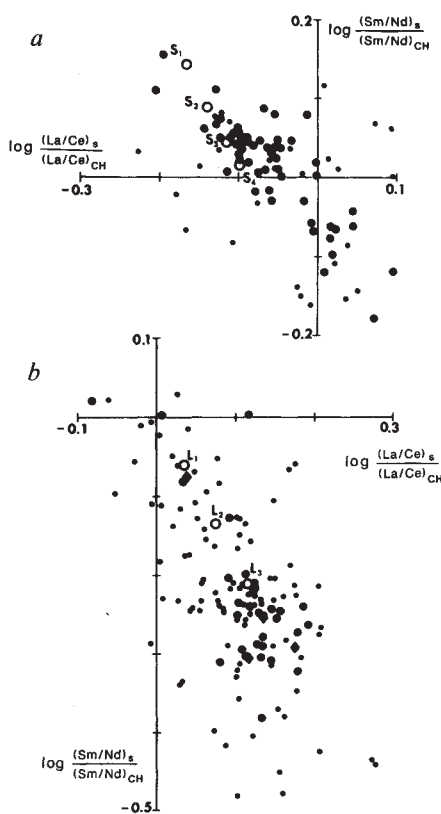


Fig. 1 Plot of epsilon Ce versus Nd (part per 10⁴ deviation from Bulk Earth isotope composition). Error bars are 2 s.d.m. within-run precision. *a*, Mostly LREE-depleted rocks. Solid circles are OIB samples³ used to calculate the best-fit mantle array, whose width is defined by average within-run Ce isotope error. Solid square is 60 Ma Skye plateau lava, relative to calculated value of 60 Ma bulk earth. Other data from ref. 1. Hatched error bars denote samples whose within-run error bars do not overlap the OIB mantle array. I = Iceland; MAR = Mid-Atlantic Ridge; EPR = East Pacific Rise. *b*, Mostly LREE-enriched rocks. Solid symbols without error bars are initial ratios². Data from refs 1, 2 and 11 and unpublished ¹⁴³Nd/¹⁴⁴Nd data: 15Z, 0.511667 ± 0.00004; 16Y, 0.510166 ± 0.00003; 7H (N.W. Jones, personal comm.), 0.510473 ± 0.000025 (all 2 SDM).

to the mantle array, and the other does not. I argue that these three Ce isotope outliers have not been sufficiently substantiated to merit a model of incoherent mantle Ce/Nd isotope evolution.

The case regarding LREE-enriched (largely continental) samples is different. Tanaka *et al.* argue that these have maintained a 'log-linear' LREE pattern throughout igneous differentiation. However, magmatic REE profiles are controlled by fractionating minerals, which are not themselves constrained to have log-linear LREE profiles¹⁰. Hence, maintenance of such patterns during crystal fractionation may be coincidental.

Ce/Nd isotope compositions of LREE-enriched rocks^{1,2} are plotted in Fig. 1b to examine empirically whether there is a linear relationship between them. Lewisian gneisses² scatter well outside error of the correlation line of Tanaka *et al.*, despite the fact that every point lies within 2σ_m limits of a La/Ce isochron² (they also lie within analytical uncertainty of the Sm/Nd isochron¹¹).

Tanaka *et al.* claim that unrelated rocks with log-linear LREE profiles should define a Ce/Nd isotope array whose slope is solely a function of the decay constants of ¹³⁸La and ¹⁴⁷Sm. This implies that the La decay constant can be determined (relative to ¹⁴⁷Sm) by isotopic analysis alone. However, this model requires that not only the samples themselves, but their mantle sources have always had log-linear LREE profiles. In contrast, Tanaka *et al.* calculated that the samples in their "mantle array" had more scattered initial isotope ratios than at present (Fig. 1b). Hence, the convergent evolution of these rocks must be a coincidence caused by small sample size, and they cannot be used to accurately constrain the La beta decay constant.

I conclude that Ce/Nd isotope systematics can be interpreted within the geochemical framework already applied to the Sr/Nd isotope diagram, and do not require special models for mantle and crustal evolution.

ALAN DICKIN

Department of Geology,
McMaster University,
1280 Main Street West,
Hamilton,
Ontario L8S 4M1, Canada

1. Tanaka, T., Shimizu, H., Kawata, Y. & Masuda, A. *Nature* **327**, 113–117 (1987).
2. Dickin, A.P. *Nature* **325**, 337–338 (1987).
3. Dickin, A.P. *Nature* **326**, 283–284 (1987).
4. Dickin, A.P., Jones, N.W., Thirlwall, M.F. & Thompson, R.N. *Contr. Miner. Petrol.* **96**, 455–464 (1987).
5. Allegre, C.J. & Turcotte, D.L. *Nature* **323**, 123–127 (1986).
6. Condomines, M., Morand, P. & Allegre, C.J. *Earth planet. Sci. Lett.* **55**, 247–256 (1981).
7. Cohen, R.S., Evensen, N.M., Hamilton, P.J. & O'Nions, R.K. *Nature* **283**, 149–153 (1980).
8. Staudigel, H. *Init. Rep. DSDP* **51**, 1331–1333 (1980).
9. Sun, S.S. *Phil. Trans. R. Soc. A* **297**, 409–445 (1980).
10. Henderson, P. in *Rare Earth Element Geochemistry* pp. 11 & 136 (Elsevier, Amsterdam, 1984).
11. Hamilton, P.J., Evensen, N.M., O'Nions, R.K. & Tarney, J. *Nature* **277**, 25–28 (1979).

TANAKA *ET AL.* reply—First, we would like to correct the term “mantle array” in the summary of ref. 1 to crustal array. We originally used the term crustal array to distinguish the clear $\epsilon_{\text{Ce}}-\epsilon_{\text{Nd}}$ array, but the term was changed in the editing of our manuscript. It was pointed out in ref. 1 that continental crustal rocks form a clearly defined ‘crustal array’, whereas mantle-derived rocks are dispersed on a $\epsilon_{\text{Ce}}-\epsilon_{\text{Nd}}$ diagram. Dickin claimed that the $\epsilon_{\text{Ce}}-\epsilon_{\text{Nd}}$ array of mantle-derived rocks (mantle array) should be sharp, as defined in Fig. 1, and that three basalts examined in ref. 1 should exist within his mantle array. We demonstrate here that (1) Ce isotopes of these basalts are not abnormal, judging from the geochemical behaviour of REE; and (2) mantle mixing is not the omnipotent to confine the isotopic values within his narrow mantle array.

The ϵ_{Ce} and ϵ_{Nd} variation in the mantle must be a result of many geochemical factors, namely La/Ce and Sm/Nd (parent/daughter) ratios, integration time of daughter nuclides, and mixing effect. Quantitative estimation of the extent of the latter two cases is difficult. Reliable rare earth element (REE) data for mantle material are few, because of alteration. Generally Mid-Ocean Ridge Basalt (MORB) is considered to have a REE pattern similar to that of mantle. Chond-

rite-normalized La/Ce-Sm/Nd diagrams for MORB and ocean island basalt (OIB) (exclude IAB) is shown in Fig. 2a. They plot with a negative correlation passing the second, third and fourth quadrants. Although MORB reflects the mantle REE pattern, partial melting and fractionation processes modify the REE pattern to be light REE (particularly La) rich. Then, the source mantle of MORB is expected to have a slightly smaller La/Ce ratio than MORB (left side in the data distribution area in Fig. 2a). S_1-S_4 in Fig. 2a are the points of model REE patterns S_1-S_4 in Fig. 2 of ref. 1. They plot just along a possible mantle La/Ce-Sm/Nd trend. S_2 and S_3 clearly represent the possible MORB source as stated in ref. 1. Mantle variation trend S_1-S_4 does not meet the origin. This is the reason why $\epsilon_{\text{Ce}}-\epsilon_{\text{Nd}}$ variation of S_1-S_4 has fan-like shape (see Fig. 3 of ref. 1). Dickin argues that a sharp mantle array stands on a mantle mixing model. Although a basic gneiss (sample 8X) in ref. 3 is not a mantle material, the 8X is a La-depleted material derived directly from a chondritic source³. Its La-depleted REE pattern (Fig. 1 of ref. 3) may be acceptable as one of the analogues of an isolated domain of depleted mantle. We examined a homogenization of isolated depleted mantle (8X) with undepleted (chondritic) mantle

and plotted the 8X sample on the top of Fig. 1b. The mixing curve definitely passes through the right-hand outside of Dickin's mantle array region.

The crustal rocks have various REE patterns with convex and concave shapes as well as a log-linear shape in their La-Sm span. Then it is natural that $\epsilon_{\text{Ce}}-\epsilon_{\text{Nd}}$ data are plotted on both sides of the $\epsilon_{\text{Ce}} = -0.112\epsilon_{\text{Nd}}$ line. Some Amitsoq gneisses are also slightly apart from the line (Shimizu *et al.*, unpublished). Chondrite-normalized La/Ce-Sm/Nd values of rocks forming continental crust (most of the rocks are from Precambrian terrane) are examined in Fig. 2b. Generally, they plot in the fourth quadrant of the diagram with negative correlation. L_1-L_3 and data for samples 16–19 of ref. 1 plot along the axial part in the whole distribution. Then, $\epsilon_{\text{Ce}} = -0.112\epsilon_{\text{Nd}}$ is considered to be an axis of the $\epsilon_{\text{Ce}}-\epsilon_{\text{Nd}}$ plot.

Interference and base line checks in our isotope measurements are already presented in refs 1 and 4. Possible interferences and peak tailing are always monitored during the measurements. On the quality of the data, we would like to emphasize that all of the data sets in ref. 1 are obtained from single solution, which provides a coherent data set among Ce and Nd isotopes and REE abundances, especially for crystalline acidic rocks.

Dickin prefers a rather smaller decay constant to $2.58 \times 10^{-12} \text{ yr}^{-1}$ on the basis of his geologic determination³. His data with trondhjemite 16Y show a rather large deviation. The trondhjemite bodies are generally intrusive to other gneiss types (ref. 5) and the REE pattern of sample 16Y is quite different from others (see Fig. 1 of ref. 3). Trondhjemite also has not been used for Sm–Nd age determination⁶. It should be excluded from La–Ce age determination. Recalculation excluding trondhjemite 16Y gives far better deviation (means square weighted deviates = 1.18). In this case, a bigger decay constant $\lambda_{\beta} = 2.68 \times 10^{-12} \text{ yr}^{-1}$ gives a coherent La–Ce age ($2,900 \pm 390 \text{ Myr}$ (2 σ)) with Sm–Nd age ($2,910 \text{ Myr}$).

T. TANAKA

Geological Survey of Japan,
Higashi 1-1-3,
Tsukuba 305,
Japan.

H. SHIMIZU
Y. KAWATA
A. MASUDA

Department of Chemistry,
University of Tokyo,
Hongo, Tokyo 113,
Japan

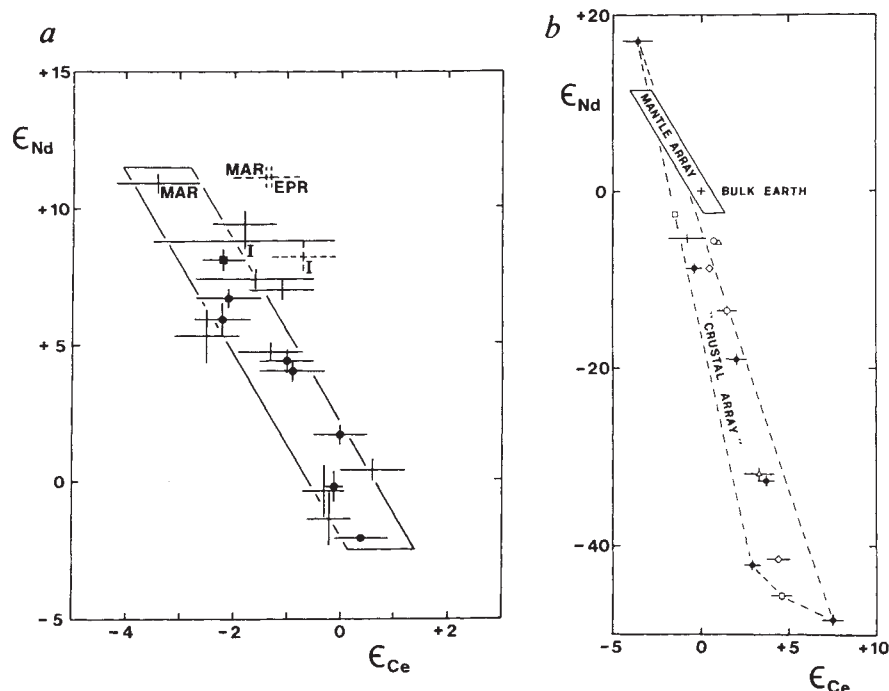


Fig. 2 $\log(\text{La}/\text{Ce})_{\text{CH}}/(\text{La}/\text{Ce})_{\text{CH}}$ versus $\log(\text{Sm}/\text{Nd})_{\text{CH}}/(\text{Sm}/\text{Nd})_{\text{CH}}$ diagrams for MORB and OIB (a), and for metamorphic and sedimentary rocks forming continental crust (b). Suffixes S and CH indicate rock sample and chondrite. $(\text{La}/\text{Ce})_{\text{CH}} = 0.377$ and $(\text{Sm}/\text{Nd})_{\text{CH}} = 0.325$ are used. These values correspond to $^{138}\text{La}/^{142}\text{Ce} = 0.00306$ and $^{147}\text{Sm}/^{144}\text{Nd} = 0.1966$. Large solid circles and small solid circles show the data obtained by isotope dilution mass spectrometry (IDMS) and by other techniques respectively. Several data are lying outside the area. Chondrite-normalized $\log(\text{Sm}/\text{Nd})_{\text{CH}}/(\text{Sm}/\text{Nd})_{\text{CH}}$ values smaller than -0.35 are the data for extremely differentiated minor rock unit. Open circles S_1-S_4 and L_1-L_3 correspond to the REE pattern of S_1-S_4 and L_1-L_3 in Fig. 2 of ref. 1, respectively. Solid diamonds show data for samples 16–19 of ref. 1. List of refs used for data points are available from the authors on request.

1. Tanaka, T., Shimizu, H., Kawata, Y. & Masuda, A. *Nature* **327**, 113–117 (1987).
2. Dickin, A. P. *Nature* **326**, 283–284 (1987).
3. Dickin, A. P. *Nature* **325**, 337–338 (1987).
4. Tanaka, T. & Masuda, A. *Nature* **300**, 515–518 (1982).
5. Weaver, B. L. & Tarney, J. *Earth planet. Sci. Lett.* **51**, 279–296 (1980).
6. Hamilton, P. J., Evensen, N. M., O'Nions, R. K. & Tarney, J. *Nature*, **277**, 25–28 (1979).

Tame thoughts on violence

Melvin Konner

Aggression: Conflict in Animals and Humans Reconsidered. By John Klama. Longman, Harlow, UK/Wiley, New York:1988. Pp.169. Hbk \$17.95; pbk £6.95.

BY THE time I got to the last page, I had racked my brains for hours trying to figure out why this book had been written. Turning back to the first lines of the preface, I found a possible clue: its genesis, reportedly, was in "the relaxed sun-and-wine ambience" of a conference in the south of France in the summer of 1980. Evidently the potent blend of photic and bibulous stimulation had so dazed the authors that they persuaded themselves that the world needed another critique of biological theories of aggression.

In the years leading up to their Provençal idyll, Ashley Montagu had published a one-sided but competent review of the case against innate aggression, Marshall Sahlins had led an anthropological assault against sociobiology, and somewhere between scores and hundreds of reviews had critically assessed books by Konrad Lorenz, E.O. Wilson, Richard Dawkins and others who had attempted to explore the intrinsic forces that might contribute to violence.

The present authors — John Durant, a zoologist at the Department for External Studies at the University of Oxford, Peter Klopfer, a zoologist at Duke University, and Susan Oyama, a psychologist at the John Jay College of Criminal Justice, City University of New York — pressed on undaunted. More critiques of behavioural biology appeared in rapid succession, threatening to scoop them. *Not In Our Genes*, by Stephen Rose, Richard Lewontin and Leon Kamin emerged in 1984. It was, like the work at hand, grotesquely biased and misleading, but at least it had some substance; it gave the earnest reviewer something to get hold of.

In 1985 Philip Kitcher published the first good (although still very biased) book criticizing sociobiology. This impressive scholarly work also failed to deflect Durant, Klopfer and Oyama from their goal: to summarize the 1980 conference and the subsequent thoughts provoked by it. They list six other authors — Erika Honore, Lisa Klopfer, Martha Klopfer, Tamara Kohn, Bryan Lessley and Nadav Nur — all of whom contributed to an unpublished volume. This the three main authors rewrote, "in order to achieve an integrity that many separate writers cannot achieve". But, oddly, they did not put their own names on the cover, preferring the invented name, John Klama. The passing years did not prevent the work from eventually coming

to light in 1988; this may constitute a record for long-term effects of a brief exposure to the environment of the south of France.

The book is a 155-page whirlwind tour through the most visible part of the ethology and sociobiology of human and animal aggression, and of human nature more generally. It consists mainly of tiny treatments — usually in two- or three-page bites — criticizing major and minor works on these subjects. With apparent ease the complex observations of Charles Darwin, Sigmund Freud, Konrad Lorenz, Paul MacLean, E.O. Wilson, Richard Dawkins, Jose Delgado, Robert Axelrod and other contributors to this debate, large and small (including, alas, myself), are disposed of. The treatments have a depth of insight and freshness worthy of examination essays ("Compare and contrast the theories of innate aggression put forth by three of the following five scientists . . ."). Textbook examples of environmental influences on behaviour round out the familiar picture. There is not a single new fact, idea, analysis or synthesis that might move forward, even a

bit, the frontier of knowledge about this grave human and scientific problem, on which the world's future may hinge; only breezy criticisms of the best efforts of others.

The scientist interested in aggression, selfishness or the idea of human nature has little to gain from this book. As for the hapless general reader, for whom it was apparently intended, even an average introductory psychology textbook has more to offer in clarifying these matters, and much else besides. Respected opponents of sociobiology! Surely you can do better than this!

Perhaps we can now solve the Klama dilemma. My guess would be that the authors read the galleys early this year, after the dazzling effect of the famed sun of Provence had finally worn off. They forebore to place their own names on the document (at this point, who could blame them?) and they had a truly creative thought: to fix the blame on the pseudonymous, eponymous John Klama!

Bad luck for a fellow who had only just been brought into existence, and who'd had no say in the matter whatsoever. I think the poor man ought to raise a Klama about it — that is, if he can find an aggressive impulse somewhere in his mythical (mystical?) hormones, experience or (perish the possibility) genes. □

Melvin Konner is Samuel Candler Dobbs Professor of Anthropology and an Associate Professor of Psychiatry at Emory University, Atlanta, Georgia 30322, USA.

Celestial eccentric

Simon Schaffer

The Peripatetic Astronomer: The Life of Charles Piazzi Smyth. By H. A. Brück and M. T. Brück. Adam Hilger: 1988. Pp. 274. £29.50, \$80.

ONE of the few unsurprising features of the life of the superb Victorian eccentric, Charles Piazzi Smyth, was his choice of career as a professional astronomer. Born in 1819 into the family of the naval officer William Smyth, whose *Cycle of Celestial Objects* (1844) became a bible for Victorian amateur observers, Charles in addition had as godfather the great Sicilian astronomer Giuseppe Piazzi.

With his father's aid, the 16-year-old Piazzi Smyth became assistant to Thomas Maclear at the Cape of Good Hope, and met John Herschel, then doyen of British astronomy. The decade in South Africa formed Smyth's future concerns. He won the patronage he needed to get the post of Astronomer Royal for Scotland, which he held from 1846 until 1888. Maclear, a precision astronomer, worried about his



Smyth — respected but troubling gadfly.

youthful assistant's interest in the optical properties of comets' tails, while Herschel aided Smyth's devotion to photography. Smyth's pictures of the Cape observatory, the authors of this welcome study suggest, may be the earliest photographs of an astronomical establishment.

The Edinburgh Observatory atop Calton Hill became Smyth's base. While conscientiously completing the pro-

gramme of reduction begun by his predecessor Thomas Henderson, and maintaining expert meteorological records, Smyth engineered challenging initiatives in physical astronomy. His achievements with the spectroscopes he bought in London and Paris or designed and adapted himself are detailed in the book. Smyth identified telluric absorption, mapped the rainband to improve weather forecasts, introduced vacuum tubes to analyse gas spectra, and investigated zodiacal and auroral light. With the physicist Alexander Herschel, John's son, Smyth established a harmonic series in the CO green band; with the Edinburgh natural philosopher Peter Guthrie Tait he proposed spectroscopic trials to detect the Earth's motion through the ether.

These programmes would have been enough to make a leader in astrophysics, but Smyth's ambitions were much wider. His research can only be understood in the context of changes in Victorian astronomy. The mid-nineteenth century witnessed a shift in disciplinary organization as the great observatories, such as Airy's at Greenwich or Struve's at Pulkovo, became centres of regimented enquiry. Airy, with whom Smyth found it increasingly difficult to maintain cordial relations, turned Greenwich into a factory-like regime, subordinating observers to the rank of menials and constructing a system of vigilant management. Smyth criticized this strategy. He denied that Adams and Leverrier, the joint 'discoverers' of Neptune, should be labelled 'astronomers', because they were not observers of the heavens but merely mathematicians. Theorists luxuriated while proper observers suffered for their vocation. His Edinburgh courses in practical astronomy were unsuccessful, and he rejected the popularization espoused by his opposite number at Glasgow, J. P. Nichol. His hopes for "communion between workers in metal and workers in paper" were frustrated. His campaign for the establishment of mountain-top observatories, initiated with a trip to Tenerife in 1856, gained some support and won him a Fellowship of the Royal Society, but it did not easily change astronomers' interests in the established metropolitan observatories.

All this made Smyth a respected but troubling gadfly on the body of his profession. More eccentric was his lengthy campaign in metrology and the interpretation of the Great Pyramid of Gizeh. Trips to Egypt and contacts with other pyramidologists convinced Smyth that the Pyramid embodied pristine mathematical truths divinely revealed to its builders. Smyth cited John Herschel's authority in his support. He fought James Clerk Maxwell over the meaning of the sacred cubit, a controversy which led to his resignation from the Royal Society. The

authors of this biography find such claims slightly embarrassing, rightly stressing the solid skills in precision measurement Smyth displayed but excusing his enthusiasm for scriptural metrology and Egyptian mysticism.

Yet Smyth's pyramidology is perhaps less inexplicable than it appears. Smyth used the relation between the sacred cubit and the inch to bolster his fight against metrication and his views on standards; any ally in the fight against French republicanism or Prussian reformist measures was warmly welcomed. Moderate theological views about precision standards were not uncommon — Maxwell himself expressed such opinions in the 1870s. But Smyth's profound political conservatism and scriptural fundamentalism uniquely combined commitment to practice over theory with a belief that the Pyramid showed the millennium would arrive in

1882. He agonized about the regimentation of his science. His travels convinced him that astronomers must change their ways, and that the ancient craftsmen to whom God had spoken provided the best model for Victorian observers. It was appropriate that the first response to his proposal for mountain observatories was developed in an 1859 Russian expedition to Mount Ararat. Only an anachronistic judgement of Victorian values prevents the recognition that the Great Pyramid could provide an excellent case for evangelical astronomy. *The Peripatetic Astronomer* provides a valuable opportunity to explore the moment at which such ancient and modern forms of worship and of science could co-exist. □

Simon Schaffer is Lecturer in the Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK.

On to experiment

Stig Stenholm

Laser Light Pressure on Atoms. By V.G. Minogin and V.S. Letokhov. *Gordon & Breach*:1987. Pp.248. \$75, £40.

LASER spectroscopy has long been concerned with the exchange of energy between electromagnetic radiation and the internal levels of atoms. When the accompanying exchange of momentum is taken into account, the theory comes into the province of the physics of laser light pressure.

Minogin and Letokhov's book is the first to present a unified and pedagogical account of this subject. It starts with a brief historical introduction, after which the text is divided into two parts. The first is a systematic and thorough account of the calculation of mechanical effects of light, the material being presented in sufficient depth to make it appropriate as the basis for a lecture course or for self-study.

The second part contains a more detailed description of the mathematical methods used in the discussion of laser cooling. The treatment is systematic and comprehensive in its coverage. It is based on an adiabatic elimination of the fast internal degrees of freedom of the atom and leads to a description of its motion in the configuration space of its centre of mass. Here the reader is expected to work a bit harder to grasp the details. The authors concentrate on the theoretical methods they have used in their own work and the theory is illustrated by experimental results. Cohen-Tannoudji's dressed-atom approach and the alternative theories worked out by Kazantsev are not discussed.

Brief accounts only are devoted to

radiation-induced diffraction in the coherent case (when relaxation is neglected) and the cooling of trapped particles, while the quantum (Lamb-Dicke) limit of traps is not covered. The text ends with a concise survey of applications, after which comes a list of 202 references. These are numbered in the order in which they appear in the text, an unfortunate practice which makes an overview difficult, and there are a few errors. But on the whole this is a representative and useful compilation, which laudably extends into 1986.

The authors have translated and revised the English edition themselves. The resulting prose generally reads very well, but there are a few annoying slips that should have been corrected in the editorial process.

The theory of laser light pressure is now well advanced and the experimental effort is growing rapidly, so the publication of Minogin's and Letokhov's volume is timely. The book will be of value for some time to come, and anyone wishing to get some insight into this subject will benefit from reading it. □

Stig Stenholm is Director of the Research Institute for Theoretical Physics, University of Helsinki, Siltavuorenpenger 20C, 00170 Helsinki, Finland.

New in paperback

- *The Anthropic Cosmological Principle* by John D. Barrow and Frank J. Tipler. Publisher is Oxford University Press, price is £9.95. For review see *Nature* 320, 315 (1986).
- *In the Company of Animals: A Study of Human-Animal Relationships* by James Serpell. Publisher is Basil Blackwell, price is £7.95, \$19.95. For review see *Nature* 323, 210 (1986).
- *The Other World: Spiritualism and Psychical Research in England, 1850-1914* by Janet Oppenheim. Publisher is Cambridge University Press, price is £12.95, \$14.95. For review see *Nature* 316, 25 (1985).

Romancing the brain

W. R. Albury

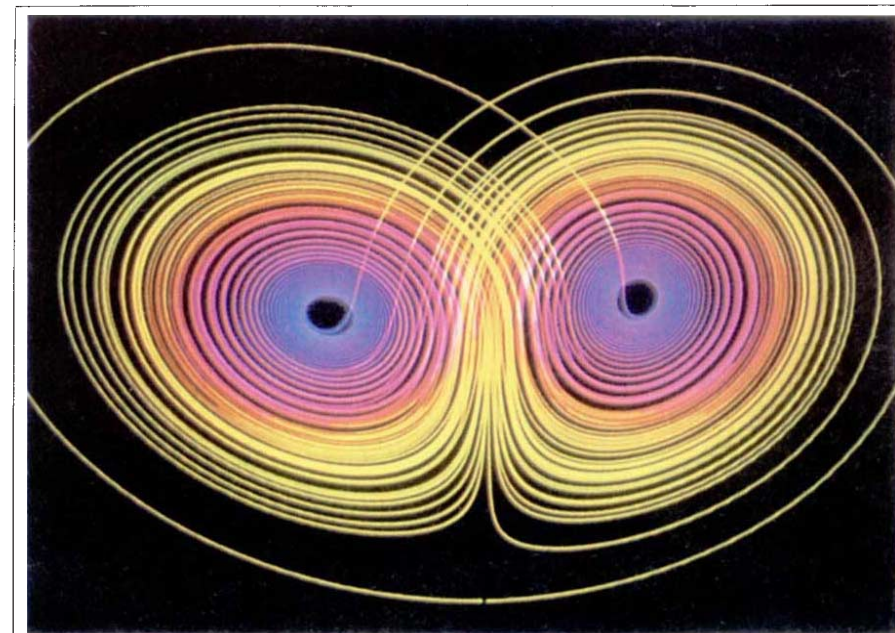
Nineteenth-Century Origins of Neuroscientific Concepts. By Edwin Clarke and L. S. Jacyna. *University of California Press*: 1987. Pp.593. \$65, £36.35.

Medicine, Mind, and the Double Brain. By Anne Harrington. *Princeton University Press*: 1987. Pp.336. \$39.50, £22.10.

ONE of the knottier problems of the history of science is to determine the extent to which past ideas, including those now considered thoroughly discredited, have contributed to shaping the present theoretical framework of scientific disciplines. The two books under review, although different in their approach and emphasis, provide interesting illustrations of this problem as it applies to the neurosciences. Clarke and Jacyna's study is the broader of the two and concentrates principally on the first half of the nineteenth century. Harrington's book is concerned with a single topic (albeit one with many ramifications) and deals chiefly with the period from 1860 to the end of the century.

Clarke and Jacyna hold that by 1850 most of the fundamental principles of modern neuroscience had been established, and in their view a key influence in the development of most of them was the romantic biology of *Naturphilosophie* in Germany and 'transcendental anatomy' in France, as well as an eclectic mixture of the two in Britain. With its stress on the unity of nature and the method of analogical reasoning, early-nineteenth-century romantic biology produced a wide range of speculative theories which later biologists would regard as wild flights of fancy. But in addition, Clarke and Jacyna suggest, this romantic philosophy of nature provided a basis for key neuroscientific ideas which have continued to be accepted by mainstream biology long after their origins had been forgotten. For example, the romantic doctrine of the unity of the organic system contributed to the extension of Marshall Hall's concept of reflex action from the spinal cord to the brain. As Clarke and Jacyna point out, however, "whereas later authors such as [Hughlings] Jackson justified their conviction in this uniformity by reference to the theory of evolution, it is important to remember that the roots of the initial breakdown of the barrier between spinal and cerebral action lay in an older biological system" (p. 147).

The authors detail the intellectual milieu of neuroscience in the early nineteenth century (including a number of other cultural tendencies besides romantic biology) and argue for a close relationship



In solving a set of nonlinear equations for convection, one would expect steady state to be reached eventually. Edward Lorenz found otherwise. Three equations with three variables described the motion of the system. Using the three variables to specify the location of a point in three-dimensional space, the evolution of the system with time can be plotted, resulting in the Lorenz attractor (above) — an image of chaos. The picture is taken from *Chaos: Making a New Science* by James Gleick, published in Britain by Heinemann on 1st June, price £12.95. For review see *Nature* 330, 293; 1987.

between scientific developments and their wider intellectual context. In addition to the concept of the reflex, the formation of ideas on the gross anatomy of the cerebrospinal axis, the histology of the nerve cell, the role of electricity in neural transmission, the localization of brain function and the special role of the autonomic nervous system are all discussed. The level of scholarship exhibited by the book is impressive, with over 200 pages being devoted to notes and bibliography. The only obvious lapse in presentation is the

• Published by Raven, shortly after the two volumes reviewed above, was Mary A.B. Brazier's *A History of Neurophysiology in the 19th Century*. The book is a sequel to her earlier work of 1984, which covered the seventeenth and eighteenth centuries.

Neurophysiology expanded dramatically in the 1800s, so the author's biographical approach becomes rather selective. Nevertheless, her choice of some 60 key figures in the discipline is exemplary, ranging from the relatively familiar (Prochaska, Magendie, Bell, Flourens, Bernard, Hall, Cajal, Jackson and Ferrier, for example) to the inappropriately remote (Eusebio Valli, Leopold Nobili, Ernst Fleischl von Marxow, Bronislav Verigo). The chapters themselves are thematically arranged around topics such as electrophysiology, cerebral localization and neuromicroscopy, and the full bibliographies and many illustrations enhance the usefulness of this well-produced volume. W.F. Bynum

failure to provide adequate diagrams to illustrate the various structural and functional aspects of the nervous system that the authors discuss.

In searching for the "origins of neuroscientific concepts", Clarke and Jacyna commit themselves to the project of evaluating nineteenth-century writings in the light of modern neuroscientific doctrines, to identify concepts from those earlier writings which appear to have retained a place in currently accepted knowledge. The problem with their historiographical approach can be illustrated by their own comments on the allegation, made in 1837, that Marshall Hall had plagiarized a book published in 1784 by Jiri Prochaska. Clarke and Jacyna explain the occurrence of this allegation, in part, "by suggesting that with the help of [Hall's] research the work of Prochaska could be viewed in a new light and thus came to appear as similar — even as an anticipation — of Hall's" (p. 119). In the absence of any explicit discussion of methodological safeguards, one cannot help wondering how many of Clarke and Jacyna's "nineteenth-century origins of neuroscientific concepts" might also be artefactual resemblances produced by using modern neuroscience to view the work of the earlier writers "in a new light". Nevertheless, whatever the doubts that one might entertain about this aspect of Clarke and Jacyna's historiography, their book represents a notable achievement of scholarly synthesis and exposition.

Harrington's study is in all senses slighter than Clarke and Jacyna's, but it

complements their book in a number of ways. The one fundamental principle of modern neuroscience that Clarke and Jacyna do not regard as having been established by 1850 is that of the localization of brain function. *Medicine, Mind, and the Double Brain* takes up this story of localization from the 1860s, with special emphasis on the question of the duality and functional asymmetry of the brain's left and right hemispheres. Harrington's project, however, is concerned not solely with the origin of these concepts but with a whole complex of late-nineteenth-century ideas on or arising from the notion of the bilateral nature of the human brain. Medical questions, and especially psychiatric ones, feature prominently, as does a consideration of the cultural context in which such ideas were formed and debated.

After a preliminary account of the situation before 1860, the book covers the investigation of and speculation about brain duality from the time of Paul Broca's work on the localization of the language faculty to the period of Hughlings Jackson's writings on hemispheric specialization some 15–20 years later. A penultimate chapter, which is really an extended footnote to the one just before it, explores Sigmund Freud's indebtedness to Jackson's concept of the double brain; while the final chapter discusses the

apparent decline of interest in brain duality after 1920, and the relatively abrupt revival of this interest in the 1960s following experiments on 'split brain' patients whose corpus callosum had been surgically severed. Having detailed in the main body of her study how the bilateral nature of the brain served in the late nineteenth century as a metaphor for society and social concerns, Harrington notes in passing a similar tendency in post-1960 writing on the subject of the brain's hemispheres.

Harrington is to be commended for including in her book a number of helpful illustrations and an extensive bibliography. Her writing style, however, must be faulted in places for its journalistic excesses, which confuse rather than enliven the discussion. Methodologically, she escapes the problem already raised in connection with Clarke and Jacyna's book by emphasizing that "the nineteenth- and twentieth-century literature on brain duality and hemisphere differences . . . are essentially discontinuous with each other" (p. 284). But she is then left with the complementary problem of the apparently independent development of

two scientific traditions with a number of common features. On this matter she offers only a few tentative speculations

The question of conceptual continuity in the sciences is just as problematical as that of conceptual discontinuity — indeed, the two questions are merely opposite sides of the same coin. Given this outcome, one is justified in asking whether the problem can be resolved, even in principle, so long as it is limited to the conceptual sphere. It seems probable that both the immediate social context of scientific research, and the broader social context in which that research is carried out, will have to be integrated coherently with the development of scientific concepts in order for a satisfactory history of science to be written. In the meantime, conceptual histories such as the two reviewed here will continue to raise fruitful questions which a more comprehensive historiography of science will someday have to answer. □

W.R. Albury is Professor of History and Philosophy of Science in the School of Science and Technology Studies, University of New South Wales, Kensington, New South Wales 2033, Australia.

Bed-time stories

Ian Oswald

Why We Sleep: The Functions of Sleep in Humans and Other Mammals. By James Horne. Oxford University Press: 1988. Pp.319. £22.50, \$49.95

JAMES HORNE is a psychologist with a primary interest in human sleep or its lack, and who describes his standpoints as heretical. The purpose of his book is the advancement of two arguments.

The first is that "we probably do not really need the last few hours of a typical night's sleep". Horne accepts that the first few hours are important for "essential repair of neurones" but proposes that the remainder of sleep is "optional", "a behaviour for occupying time". The proposition rests heavily on studies with healthy young volunteers who get by without currently measurable defects when repeatedly short of sleep. In discussing the best-known American research of this kind, Horne omits the descriptions of fatigue, with general conservation of energy, easy discouragement and dampening of emotional involvement. He remarks that his own volunteers did not make use of their extra waking hours, but did more "wasting time". Horne's opinion owes nothing to the clinic, the elderly or the sick. Finally, he turns about, says he is not advocating reducing sleep to six hours, and adds a summary of one of the large studies that have now demonstrated

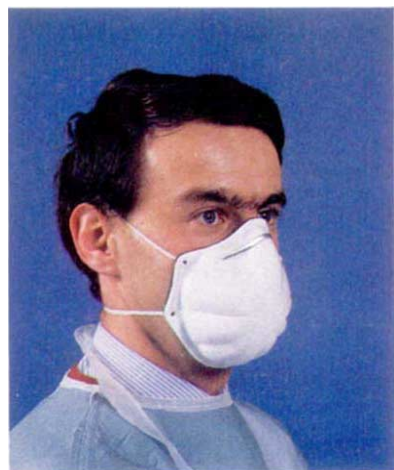
that people who believe they habitually sleep six hours or less show a higher rate of mortality from all causes in subsequent years.

Horne's belief in optional sleep also rests on the fact that sleep-deprived people, when catching up, never take as many extra hours of sleep as they originally lost. Sleep may have intensity of restitutive function, not just duration, but this is something upon which he does not dwell.

The second argument set forth in the book is that although sleep is indeed associated with restitution of all the body's tissues in rodents, this is positively not so for human beings. For Horne, only the brain can be repaired during human sleep. The time of the largest and slowest electrical brain waves during sleep is that which Horne would see as bringing restitution of brain: it is, as he says, the time when brain metabolism is lowest. In his wish to aver that restitution will not be taking place in other human tissues at this time, he says it cannot be taking place *because* the body's metabolism is at its lowest, which to me seems to be inconsistent. He does not mention that others have argued that it is because metabolism is low, rest maximal and degradation minimal, that net repair is most possible. Horne's arguments failed to persuade me that human beings should not be on a continuum with other animals. □

Ian Oswald is a Professor in the Department of Psychiatry of the University of Edinburgh, Royal Edinburgh Hospital, Morningside Park, Edinburgh EH10 5HF, UK.

Lightweight CHEMISORB Respirators with activated-carbon + particle filters



The CHEMISORB 'Super'-Masks protect from chemical carcinogens, organic vapours, toxic gases, odours, mists, fumes & smoke.

SAFELAB SYSTEMS LTD.,
Bush House, 72, Prince Street,
Bristol BS1 4HU, Great Britain
Phone and Fax: 0272-394455

The second law of thermodynamics: entropy, irreversibility and dynamics

Peter V. Coveney*

Theoretical Chemistry Department, Oxford University, Oxford OX13 3PG, UK and
Faculté des Sciences, CP 231, Campus Plaine, Université Libre de Bruxelles, Bruxelles, Belgium

Irreversibility is at once a profound and an elusive concept. We are all aware of the directed, irreversible arrow of time which dominates our own existence. Until recently, however, it had been thought that the fundamental dynamical laws were incompatible with the objective existence of the irreversible processes that thermodynamics purports to describe.

MOST scientists accept the existence of irreversible processes as a brute fact of life. Being macroscopic, their definition may be found in thermodynamics, which is concerned with the formalization of the facts of everyday experience. A reversible process is one in which a given system may be returned to its initial state, even after the performance of work and the exchange of heat, with no accompanying effects arising in the surroundings. All other processes are irreversible; for example, even if it is possible to return a system to its initial state, changes will occur in the surroundings (transfer of heat, and so on). Clearly, reversible processes are a very special case.

The second law

The distinction is sharpened by the second law of thermodynamics, which asserts that, for any irreversible process, the function S , known as the entropy, must increase, whereas for a reversible process it remains constant. More particularly, the entropy of an isolated system increases monotonically until it reaches its maximum at equilibrium—a form of words that implies a sense of the direction of time, as does the familiar statement of the second law in the form

$$dS \geq 0 \quad (1)$$

where the equality applies only to reversible processes; more exactly, this should be written

$$\frac{dS}{dt} \geq 0 \quad (2)$$

This makes it clear that the positive direction of time (t) is linked to the increase in S during an irreversible process. Indeed, the second law of thermodynamics is the only law of physics that introduces a sense of direction in time.

Thermodynamics of irreversible processes

Because of the very general, yet vague, nature of the second law, most of its applications during the first century after its formulation 130 years ago were restricted to the special case of equilibrium, when the equality sign of equation (2) applies. At equilibrium, other thermodynamic potentials can be used instead of the entropy itself, such as the Helmholtz (A) or Gibbs (G) free energy, which locate the position of equilibrium for appropriate boundary conditions (fixed temperature and volume, or fixed temperature and pressure for A and G respectively).

One tends to overlook the vast range of phenomena that comprise the field of non-equilibrium thermodynamics, the irreversible processes that drive macroscopic systems to thermodynamic equilibrium. At first sight, the so-called 'heat death' of the Universe (first articulated by Helmholtz¹ and popularized

later by Eddington²) predicted on the basis of the second law, appears to support the notion that the low entropy order we perceive around us is gradually, yet inexorably, degenerating into chaos as the Universe tends towards thermodynamic equilibrium.

Accordingly, it is a remarkable result of the past few decades of research in non-equilibrium thermodynamics that non-equilibrium or irreversible processes may be a source of order within a system.

Order out of chaos

In an open system capable of exchanging both energy and matter with the surroundings (Fig. 1), entropy flow, $d_e S$, may be distinguished from entropy production, $d_i S$. The former is the entropy exchanged with the surroundings, and the latter the entropy produced internally by the system. The second law asserts that entropy production satisfies the inequality

$$P \equiv \frac{d_i S}{dt} \geq 0 \quad (3)$$

(with the signs taken as before). The thermodynamic potentials (such as A and G) exist only for the special case of equilibrium, because now the inequality does not involve the total differential of the entropy S . If one seeks to go beyond equilibrium, more needs to be said about the nature of the entropy production. By the approach of Prigogine, which has come to be known under the name of generalized thermodynamics, one makes the assumption of 'local equilibrium', that out of equilibrium the entropy still depends on the same macroscopic variables as those at equilibrium. In this way we arrive at the basic equation of non-equilibrium thermodynamics

$$P \equiv \frac{d_i S}{dt} = \sum_k J_k X_k \geq 0 \quad (4)$$

where J_k is the flux or rate and X_k the generalized force of the k th irreversible out-of-equilibrium process.

Equation (4) generalizes an expression derived by the Belgian thermodynamicist de Donder, founder of the 'Brussels School', relating the entropy production during a chemical reaction to the product of affinity and the reaction rate³.

In linear non-equilibrium thermodynamics, where the deviations from equilibrium are small, the fluxes J_k depend linearly on the forces X_k as

$$J_k = \sum_{k'} L_{kk'} X_{k'} \quad (5)$$

With the phenomenological coefficients $L_{kk'}$ (equal to $L_{k'k}$ by Onsager's reciprocity relation³) taken to depend only on the equilibrium parameters (a stronger requirement than linearity), the entropy production $P = d_i S/dt$ itself functions as a potential, in the sense that the evolution settles down to a non-equilibrium steady-state which minimizes the dissipation⁴.

* Present address: Department of Chemistry, University College of North Wales, Bangor, Gwynedd LL57 2UW, UK.

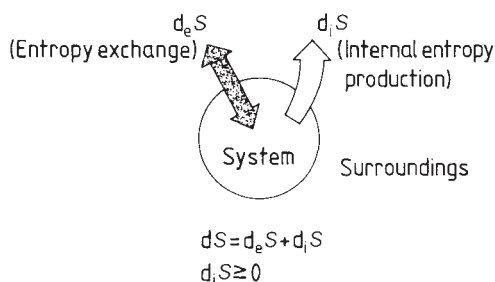


Fig. 1 Entropy changes occurring in an open system.

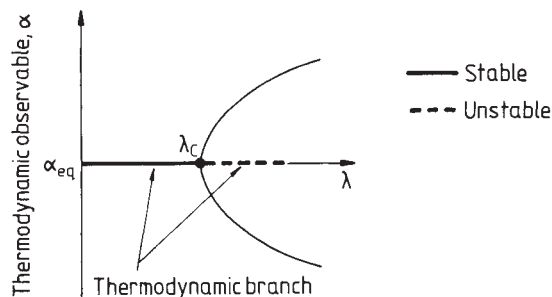


Fig. 2 Schematic bifurcation diagram indicating how the thermodynamic branch of minimum entropy production becomes unstable beyond a certain value, λ_c , of the parameter λ that measures the distance of the system from thermodynamic equilibrium. The dissipative structures are associated with the evolution of the system beyond the point λ_c .

In the non-linear regime, far from equilibrium, it is, however, no longer possible to describe the system in terms of a scalar potential. Thermodynamic stability analysis⁵ shows that the second-order change in entropy about the stationary value, $\delta^2 S$, satisfies the equation

$$\frac{1}{2} \frac{\partial}{\partial t} (\delta^2 S) = \sum_k \delta J_k \delta X_k \quad (6)$$

where δJ_k , δX_k are the deviations of J_k , X_k from their stationary values.

The important point is that the right-hand side of equation (6) now has in general no particular sign. Depending on the detailed form of the forces and fluxes, the thermodynamic branch may become unstable (see the bifurcation diagram in Fig. 2): the state of minimum entropy production is no longer the state to which the system evolves. Instead, we can then expect to find what have been termed by Prigogine dissipative structures, ordered states of matter occurring beyond the bifurcation point at which the thermodynamic branch becomes unstable. Their existence stems from the exchange of energy and matter with the environment, together with the production of entropy (dissipation) by the system.

These non-equilibrium dissipative structures abound in chemistry and biology. Chemical reactions form the best example of irreversible systems to which such analyses may be applied; a necessary (but not sufficient) condition for their existence is that the chemical (or biochemical) reaction scheme should involve an autocatalytic step⁶. The J_k and X_k terms in equations (4) and (6) are then the rates of reaction and the affinities of the species present (and possibly also the diffusive flux and chemical potential gradients if spatial inhomogeneities are relevant). Over the past two decades, many complex chemical reactions have been discovered whose model kinetic schemes possess the required non-linearities and thus manifest

so-called self-organization under far-from-equilibrium conditions.

Dissipative structures, such as chemical clocks and moving or stationary chemical waves, involving a spatially inhomogeneous distribution of matter with or without time dependence, can all be found in different types of reaction systems far from equilibrium. The classic example in chemistry is the Belousov-Zhabotinskii reaction, which involves a mixture of bromate ion, malonic acid and cerium(IV) ions dissolved in citric acid solution (or some variant on this theme)^{6,7}. Thanks to Noyes and Field⁷ and others, the essential features of the reaction mechanism are known; kinetic models confirm that the oscillations of reaction-intermediate concentrations are those obtained experimentally (Fig. 3 and ref. 8).

This example recalls the glycolytic oscillations in the glycolytic cycle, the source of adenosine triphosphate in living cells, which was the first confirmed example of a biological dissipative structure⁹. Because biochemical reactions usually occur far from equilibrium, and hence autocatalytic activity is commonplace, the importance of such oscillations for the control of regulatory processes in biological systems is readily appreciated.

Statistical mechanics

Although at equilibrium and nearby, the entropy or related thermodynamic potentials such as the free energy are of central importance, for a system far from equilibrium, such potentials do not exist: the vagueness of the second law leaves open the question of the precise meaning of the entropy out of equilibrium. This is undoubtedly why so much of thermodynamics has been restricted to equilibrium conditions, but the complexity that may arise far from equilibrium requires an investigation of the more general nature of entropy.

Can more light be shed on the problem by statistical mechanics? In equilibrium thermodynamics, the macroscopic level is rendered understandable by a microscopic description that takes account of the atomic and molecular constitution of the underlying matter. Equilibrium statistical mechanics is usually formulated in terms of the partition function Z (ref. 10):

$$Z = \sum_i \exp(-\beta E_i) \quad (7)$$

where $\beta = 1/kT$, T being the temperature, k Boltzmann's constant and E_i the energy levels of the system, which are in principle known from the laws of quantum mechanics that govern the behaviour of matter on the microscopic level. Given the partition function, all the equilibrium properties of matter can be calculated. Of course, the partition function will be known exactly only for a few very simple systems, primarily involving non-interacting particles (ideal systems), but this is merely a practical difficulty, as approximations can always be invoked in other cases.

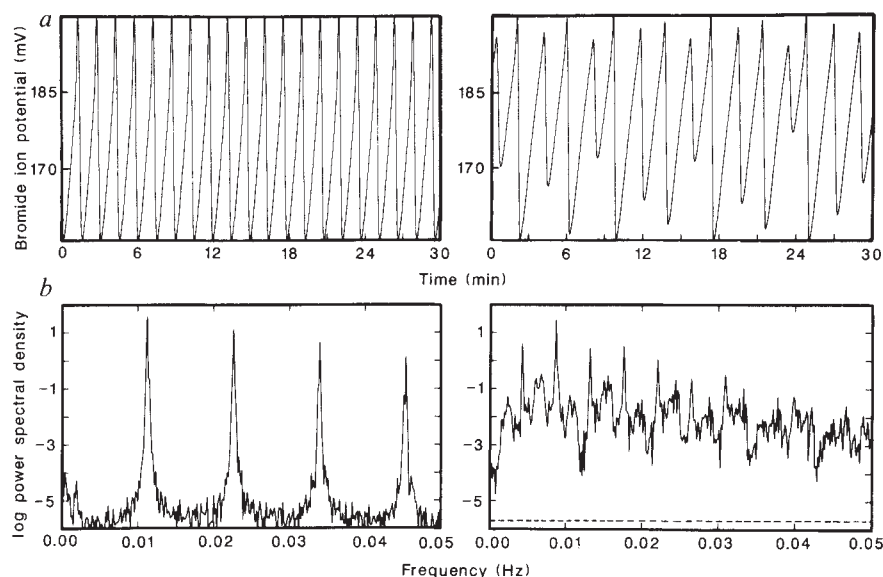
Thus, for a closed system able to exchange energy but not matter with the environment, the microscopic version of the equilibrium entropy is given by

$$S = k \ln Z + kT \left(\frac{\partial \ln Z}{\partial T} \right)_v \quad (8)$$

In non-equilibrium statistical mechanics, we might similarly expect to be able to furnish a microscopic interpretation of irreversible dynamical behaviour, or the irreversible time evolution of macroscopic systems. What, then, is the general formulation of the entropy out of equilibrium?

Although the partition function cannot be defined out of equilibrium, for the dynamical description of a large number ($\sim 10^{24}$) of particles it is in any case more informative to work with the distribution function for classical systems, or the density operator for quantum systems (both denoted by ρ). These quantities represent the probability of finding the system in a given state in phase space (the space of all the coordinates $\mathbf{q}$ and momenta $\mathbf{p}$ of the particles in the system in a classical descrip-

Fig. 3 Data for periodic and non-periodic states observed in the Belousov-Zhabotinskii reaction. *a*, Bromide ion potential time series, *b*, Power spectral density. In the non-periodic state, the broadband noise is more than three orders of magnitude above the instrumental noise level, which is indicated by the horizontal dashed line⁸.



tion) or in Hilbert space (the space of states, or wavefunctions, in quantum theory).

Classically, there is no problem. The dimensionality of phase space is $2DN$, where N is the number of particles and D the dimensionality of the underlying geometric space. If the position of this point in phase space is exactly known, its subsequent motion will be calculable from Newton's laws; if there is some uncertainty, there will be a similarly calculable flux of probability. In quantum mechanics, the dimensionality of the space, that of the space of states, is usually infinite, but that makes no essential difference; the point that represents the state of this system moves, tracing the evolution of the system as it does so. Formally, in the Liouville spaces in which ρ is defined, in both classical and quantum theory, ρ satisfies the Liouville equation

$$i \frac{\partial}{\partial t} \rho_t = L \rho_t \quad (9)$$

where t is time and the Liouville operator (or Liouvillian) is given in terms of the Hamiltonian H (the total energy of the system) by

$$L = \begin{cases} i\{H, \cdot\} \equiv i \left(\frac{\partial H}{\partial \mathbf{p}} \cdot \frac{\partial}{\partial \mathbf{q}} - \frac{\partial H}{\partial \mathbf{q}} \cdot \frac{\partial}{\partial \mathbf{p}} \right) \\ [H, \cdot] \equiv H \times 1 - 1 \times H \end{cases} \quad (10)$$

the braces denoting the Poisson bracket in the classical case, and the square brackets the commutator in the quantum case¹⁰. In quantum mechanics, operators such as L which act on the density operator are called superoperators, and Liouville space is known as superspace as it is the direct product of Hilbert space with itself. One should note that the formal appearance of the equations of motion is, however, the same for the two cases in this formulation of dynamics. The Liouville space formalism provides a means of extending the treatment of statistical mechanics from equilibrium (when $\rho_t = \rho_{eq}$ is independent of time) to more general non-equilibrium situations (when ρ_t is dependent on the time)¹⁰.

But there is a serious problem with the Liouville equation (9): as it is left invariant under time reversal (when $t \rightarrow -t$ and $L \rightarrow -L$, the latter because the momenta $\mathbf{p}$ are then reversed), it can only describe reversible processes. This time-reversal symmetry is a fundamental feature of essentially all dynamical equations. For example, at equilibrium, the Gibbs expression for the entropy can be written in terms of ρ_{eq} as¹⁰

$$S_G^{eq} = \begin{cases} -k \int_{\Gamma} d\mu \rho_{eq} \ln \rho_{eq} & \text{(classical)} \\ -k \text{Tr} \rho_{eq} \ln \rho_{eq} & \text{(quantal)} \end{cases} \quad (11)$$

where $d\mu$ is the invariant measure (volume element) of the phase space Γ , and Tr denotes the trace (the sum of the diagonal elements) of the matrix product which follows. This relation corresponds exactly to that given in equation (8) for the equilibrium entropy, because $\rho_{eq} = e^{-\beta H} / Z$.

But such an expression cannot represent the entropy of a system out of equilibrium, because we may use equation (9) to show that

$$\frac{\partial S_G}{\partial t} = -k \frac{\partial}{\partial t} \int_{\Gamma} d\mu \rho_t \ln \rho_t = 0 \quad (12)$$

which is to say that S is a constant of motion, in contradiction with the second law. Thus for an isolated system, where L is time-independent, the solution to equation (9) for ρ_t at time t in terms of the initial condition ρ_0 at time $t=0$ is given by

$$\rho_t = \exp(-iLt) \rho_0 \quad (13)$$

and defines an evolution (super-)operator $U_t = \exp(-iLt)$. This is a one parameter (t) unitary group, implying conservation of probability, and hence isentropic (constant entropy) evolution. As long as the dynamical evolution is unitary, irreversibility cannot arise. This is the fundamental problem of non-equilibrium statistical mechanics. Several attempts have been made to solve it, including most notably Boltzmann's $\mathcal{H}$ -theorem, the so-called coarse-graining technique, and more recently the non-unitary transformation theory associated with the Brussels School³³.

Boltzmann's $\mathcal{H}$ -theorem. Boltzmann's contribution to the question of a dynamical interpretation of the entropy remains as a landmark more than one hundred years later¹¹. By approximating the exact dynamical equations (equation (9)) directly in terms of a kinetic equation for the velocity distribution function, $f(v_1) \equiv f_1$, of a molecule in a gas, he obtained the equation

$$\frac{\partial f_1}{\partial t} + v_1 \frac{\partial f_1}{\partial x_1} = \iint d\Omega dv_1 \sigma \{f_1' f_2' - f_1 f_2\} \quad (14)$$

where σ is the collision cross-section, $d\Omega$ is the element of solid angle, and the primes refer to the new velocities following two-body collision. The essential approximation was the replacement of the two-particle reduced distribution function

by a product of two single-particle functions—the assumption of ‘molecular chaos’—which takes the particles as being uncorrelated before collision.

Using equation (14), Boltzmann then showed that the quantity $\mathcal{H}$, defined by

$$\mathcal{H}(t) = \int_{\Gamma} d\mu_1 f_1 \log_e f_1 \quad (15)$$

satisfies the inequality

$$\frac{\partial \mathcal{H}}{\partial t} \leq 0$$

Then the non-equilibrium Boltzmann entropy can be identified with

$$S_B(t) = -k\mathcal{H}(t) \quad (16)$$

Note the similarity with the Gibbs entropy, equation (11); the difference is that the former depends on a reduced distribution function, f_1 .

The Boltzmann equation (14) is a good approximation for the description of dilute gases (where probably only binary collisions occur), but it cannot be extended in any simple manner to more general situations. Despite all the progress that has been made in employing kinetic theory for treating transport phenomena (viscosity, diffusion and thermal conductivity) in more complicated cases¹⁰, it has not proved possible to generalize the $\mathcal{H}$ -theorem itself by conventional means.

Coarse-graining. Another quite popular approach has been to relegate the whole question of irreversibility as illusory. The reversible dynamical laws of the microscopic domain are regarded as unassailable in view of the tremendous successes of classical and quantum mechanics in many fields, and the irreversibility manifested on the macroscopic level apparently arises from the various approximations made in the course of observation. Irreversibility is admitted into the description by asserting that we only observe a coarse-grained probability; in other words we can discern only finite, rather than infinitesimal, regions of phase-space. Coarse-graining leads to apparent irreversibility at the macroscopic level, while preserving reversible evolution in the fundamental dynamical equations (see Fig. 4). Entropy then acquires a rather subjective aspect, reflecting only our loss of information about the details of a dynamics that is fundamentally reversible¹².

But if the case of classical statistical mechanics is considered, the traditional coarse-graining arguments fail to differentiate properly between the different asymptotic behaviour of all the trajectories considered in an ensemble (bundle of trajectories), which may become distinct at later times. This leads to inconsistencies in the formulation that have never been satisfactorily overcome.

Of course, should the coarse-graining argument be correct, it would flaw the whole programme adopted by Boltzmann; also, we would be forced to accept that macroscopic processes as manifestly irreversible as the functioning of our brains and life itself are really only the result of our own approximations. This paradox is summarized by the statement ‘Irreversibility is either true on all levels or on none: it cannot emerge as if out of nothing, on going from one level to another’¹³.

One of several possible approaches³³ demanding a proper place for irreversibility on the microscopic level has been made over the past twenty years or so by the Brussels School. The whole subject is rather revolutionary in nature and is still in a state of evolution. The following account is not chronological, but rather presents an overall perspective of this work from today’s standpoint.

Classical systems

Although quantum mechanics is taken to be a more correct microscopic theory of dynamics, there is much to be learned from a consideration of the situation in classical mechanics. This is because irreversibility turns out to be closely related to

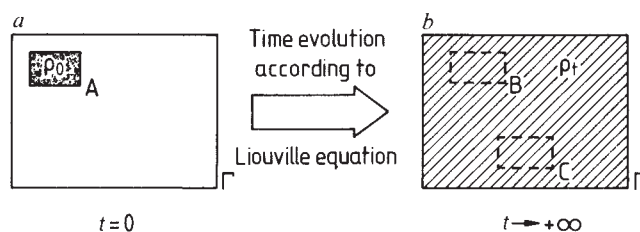


Fig. 4 *a*, System prepared with probability distribution ρ_0 confined to region A of phase space Γ . *b*, For appropriate dynamical systems, as $t \rightarrow +\infty$, any finite, local region in Γ , such as B or C, looks like the entire phase space. Our loss of information as to where the system is localized is the origin of coarse-grained irreversibility.

dynamical instability, the mechanism for which is well understood in the latter case.

Classical dynamics The simplest way of attempting to incorporate the second law into classical dynamics would be to construct a function with the properties of entropy as a monotonically increasing function of time, already discussed above in connection with Boltzmann’s $\mathcal{H}$ -theorem. Such a function is known as a Lyapounov variable or functional. But this attempt must be still-born by virtue of Poincaré’s theorem which rules out the existence of any such function on phase space¹⁴: the reason is basically that given enough time, any system will return (arbitrarily close) to its initial state in phase space (even if the ‘recurrence time’ is many times the age of the Universe).

But progress in the theory of classical dynamical systems has shown that this is not the end of the story. A full account of the classification of dynamical systems would take us deep into the highly mathematical realm of ergodic theory¹⁵, but this is not necessary to present the essential ideas. The key point is the existence of varying degrees of instability of motion. By instability we mean sensitivity of the long-time behaviour of the system to the initial conditions. Consider a system with Hamiltonian (total energy) H and initial conditions given by the point in phase space $\{q(0), p(0)\}$, where q denotes the coordinates and p the momenta of the particles in the system; then the future (and past) evolution (or trajectory) of the system can be predicted deterministically from Hamilton’s equations of motion

$$\frac{d}{dt} q(t) = \frac{\partial H}{\partial p}, \quad \frac{d}{dt} p(t) = -\frac{\partial H}{\partial q}$$

If the system is stable, a slight change in the initial conditions will not alter its long-time behaviour. For increasing degrees of instability, closely neighbouring initial states will tend to have increasingly different long-time evolution. For highly unstable dynamical systems, such as those termed *K*-flows in ergodic theory¹⁵, neighbouring trajectories diverge exponentially in time. (Indeed, a rigorous analysis of such systems demands that we leave the Hamiltonian formalism, and instead work with what are generally referred to as abstract dynamical systems.) A number of idealized systems of great physical interest (such as the infinite ideal gas, the hard sphere gas, motion of three billiard balls on a table) are known to be *K*-flows. Now, if dynamics is to be capable of describing what we observe, then we are obliged to take into account the fact that it is quite impossible for us to achieve absolute precision (infinite information) in measuring the initial conditions. This assertion, once accepted, has far-reaching implications. Because the passage from finite regions to points in phase space now becomes a singular limit, the whole concept of trajectories, which has been the basis of classical mechanics for three hundred years, becomes no longer operationally meaningful for a class of sufficiently unstable dynamical systems. This, in turn, forces us to reconsider the question of determinism even in classical mechanics.

One concludes that for systems displaying sufficiently strong dynamical instabilities, the dynamics must instead be formulated

in terms of distribution functions on phase space. This is an interesting result, for we see that it brings us back to the kind of description used in non-equilibrium statistical mechanics. But the Liouville space formalism enables us to broaden the definition of a dynamical variable: in addition to simple multiplication by phase space functions, we can now introduce more general operators that act on the distribution function ρ . In particular, the Liouville operator L (equation (9)), being the Poisson bracket of the Hamiltonian, contains derivatives with respect to both $\mathbf{q}$ and $\mathbf{p}$. Without violating Poincaré's theorem, we can now seek Lyapounov variables of the form

$$\int_{\Gamma} d\mu \rho_t M \rho_t,$$

where M is now an appropriate dynamical operator playing the role of entropy on the microscopic level.

What are the conditions for the operators M to exist? As we might anticipate, not all dynamical systems admit the existence of variables M : a minimum degree of instability is essential. For ergodic systems, whose trajectories 'fill' the entire phase space compatible with the conservation of energy, the conditions can be stated very precisely in terms of the spectral properties (the nature of the eigenvalues) of the Liouville operator: the Liouvillian must possess a continuous spectrum along the entire real line¹⁶.

There is a class of ergodic dynamical system that fulfills this condition, and for which Lyapounov variables exist. This is the set of K -flows already mentioned above. A number of remarkable features emerge for these systems. One can define an internal time operator T which has the commutation property

$$i[L, T] = i(LT - TL) = 1 \quad (17)$$

and reflects the age of the system. An entropy operator M can then be constructed as a monotonically decreasing positive operator function of T ¹⁷. In this way, a meaning, has been given to the entropy entirely within the framework of the dynamical description.

There is another class of classical system, which is neither integrable (in the sense that analytic action variables do not exist) nor ergodic, but for which the same kind of dynamical instabilities exist as in the latter case. These non-integrable systems, which begin with the famous three-body problem, are of major importance in physics. The construction of the entropy operator here involves the use of subdynamics, discussed further below. Work is still going on to clarify aspects of this more complex case, but it is important to remark that the source of irreversibility in Boltzmann's equation stems from non-integrability rather than from ergodic properties of the dynamics.

A very interesting feature of equation (17) is the non-commutativity of the time operator T with the Liouvillian L . One is immediately led to a comparison with the situation in quantum mechanics, where the non-commutativity of observables gave rise to Bohr's notion of complementarity¹⁸. By analogy, we can say that the dynamical and thermodynamic descriptions of such unstable systems are complementary: a complete certainty in the dynamical description in terms of trajectories (through L) renders thermodynamics (through T) meaningless, whereas, conversely, a fully thermodynamic description is incompatible with a complete knowledge of the dynamics^{16,19}.

Non-unitary transformation theory. We can now define a non-unitary transformation Λ through

$$\Lambda = M^{1/2} \quad (18)$$

in terms of which the reversible unitary group U_t is related to an irreversible semigroup W_t^* by the relation

$$W_t^* = \Lambda U_t \Lambda^{-1} \quad (t \geq 0) \quad (19)$$

The essential point about a semigroup for present purposes is that it satisfies the equality

$$W_t^* W_s^* = W_{t+s}^* \quad (20)$$

only for $t, s \geq 0$, whereas the corresponding group property

$$U_t U_s = U_{t+s} \quad (21)$$

is true for all times t, s . The non-unitary transformation, equation (18), therefore breaks the time-symmetry of the evolution originally present in equation (9) and the process becomes irreversible. The transformed distribution function $\Lambda\rho$, known as the physical representation ${}^p\rho$, causes the originally deterministic evolution

$$\rho_t = U_t \rho_0 \quad (22)$$

to be transformed, using equation (19), into a stochastic Markov process:

$${}^p\rho_t = \Lambda\rho_t = \Lambda U_t \rho_0 = W_t^* \Lambda\rho_0 = W_t^* {}^p\rho_0 \quad (23)$$

Then any convex functionals of ${}^p\rho$, such as

$$\int_{\Gamma} d\mu ({}^p\rho_t)^2 \quad \text{or} \quad \int_{\Gamma} d\mu {}^p\rho_t \ln {}^p\rho_t$$

will satisfy an $\mathcal{H}$ -theorem with no approximation^{16,17}.

Quantum systems

When quantum mechanics is considered it is much more difficult to obtain rigorous results, because 'instability' in quantum systems is not yet properly understood. Nevertheless, a similar line of reasoning to that just described for the classical case has already led to some rather general conclusions²⁰.

Again, there are fundamental reasons, which follow along qualitatively similar lines to those mentioned above, that necessitate a superspace formalism to incorporate the possibility of defining an entropy superoperator M . Indeed, it can be shown that such a superoperator cannot exist if the Hamiltonian (energy) operator has either a purely discrete or a bounded spectrum²⁰. It follows that no entropy superoperator can be defined for finite quantum systems, in marked contrast to the situation in classical mechanics. Conversely, when the Hamiltonian has a spectrum that in a limiting sense becomes continuous from 0 to $+\infty$, then a time superoperator exists and hence so too does M . Thus, we should look at large quantum systems in the thermodynamic limit, corresponding to the limit of an infinite number of particles in an infinite volume, such that the density or concentration of the particles is a finite constant. This could be just the right kind of system to display irreversibility.

The physical meaning of the entropy superoperator—when it exists—is, however, similar to that in classical mechanics. Through the introduction of the non-unitary transformation of equation (18), we are led to a semigroup description of the time evolution in terms of an irreversible quantum stochastic process.

On a fundamental level, the existence of the entropy superoperator is of great significance with respect to one of the major difficulties in quantum theory, the measurement problem. According to the conventional formulation of the theory as defined by von Neumann²¹, there are two distinct types of transformation which occur to the wavefunction Ψ , the quantity containing all the information to be known about a given system. On the one hand we have the unitary (reversible) deterministic evolution of pure states Ψ , governed by the Schrödinger equation

$$i \frac{\partial}{\partial t} \Psi_t = H \Psi_t \quad (24)$$

(H is the Hamiltonian of the system); on the other hand, we have the supposed reduction or collapse of the wavefunction, which occurs irreversibly and non-deterministically when a measurement is made on the system, yet for which no mechanism is provided. The pure states remain pure states as time progresses and are converted into mixtures—for which a density operator formalism is required—by the measurement process.

However, the entropy superoperator M introduced above can be shown to be non-factorizable²⁰. This means that it cannot be

written as a product of Hilbert space operators and, as a result, it transforms pure states into mixtures. Put another way, if the system evolves irreversibly (that is, if M exists), there is no longer any reason to be restricted to the notion of the wavefunction. Conceptually, this is the quantal equivalent of the classical case where, when M exists, trajectories must be replaced by distribution functions. Thus we have in principle a resolution of the measurement problem in quantum mechanics, because we have now consistently taken into account both the macroscopic nature of the measuring apparatus, and the irreversible nature of the measurement act itself, the two fundamental aspects of measurement (as repeatedly stressed by Bohr²²).

Irreversibility in real systems: subdynamics

The results described above, although of great value, are too formal to be useful in describing irreversible processes in real systems. If we do not depart from the idea that the 'ultimate' microscopic description is actually quantum-mechanical, it follows from the foregoing considerations that we must study the behaviour of large many-body systems in the thermodynamic limit if we are to be able to introduce an entropy superoperator.

The subdynamics theory of irreversible processes^{23,35}, another development of the Brussels School, is both more physical and more suitable for performing such a task. Aside from furnishing the first clear cut description of the relation between dynamics and thermodynamics, it has proved to be a very powerful method for deriving kinetic equations in non-equilibrium statistical mechanics¹⁰. Here too it had already been noted that a super-space formalism, together with the spectrum of L becoming continuous, are both essential for the description of dissipative (that is, irreversible) processes^{23,24}.

The subdynamics theory is rather involved but, briefly, one begins with the usual perturbation expression for the true system

$$L = L_0 + \delta L \quad (25)$$

where L_0 represents a solvable problem, and δL are the interactions between the units defined by L_0 . The approach is based on the introduction of a complete set of correlation states ν , (eigenstates of L_0) whose hermitian projection operators $P^{(\nu)}$ commute with L_0 . For the case of a dissipative system in the thermodynamic limit, we can no longer diagonalize the full Liouvillian superoperator L , but we can show that a new complete set of non-hermitian projection operators $\Pi^{(\nu)}$ exists, each one of which commutes with L :

$$[\Pi^{(\nu)}, L] = \Pi^{(\nu)} L - L \Pi^{(\nu)} = 0$$

If we define the components of the density operator by $\rho^{(\nu)} \equiv \Pi^{(\nu)} \rho$, then we see that each component evolves independently

$$i \frac{\partial}{\partial t} \rho^{(\nu)} = L \rho^{(\nu)}$$

hence the origin of the term subdynamics. A generalization of the conventional unitary quantum-mechanical transformation theory can now be defined that relates the projection operators $\Pi^{(\nu)}$ and $P^{(\nu)}$ through

$$\Pi^{(\nu)} = \Lambda P^{(\nu)} \Lambda^{-1} \quad (26)$$

This non-unitary transformation is essentially the same as that described previously, although it contains detailed information on the physical nature of the system concerned; as before, it converts the unitary description into the physical representation, where once again the evolution possesses an arrow of time

$${}^p \rho_t = W_t^{*p} \rho_0 \quad (27)$$

The subdynamics formalism has very recently been extended to include the interaction of a large system with time-dependent external fields^{25,26}. This should enable the theory to tackle the important problem of irreversibility in matter-radiation interactions³⁰, as well as other problems of major interest in statistical

mechanics involving irreversible evolution of macroscopic systems in the presence of external fields³⁴.

Conclusions

This article has presented the case for a more unified description of phenomena, in which the second law of thermodynamics is built into the dynamical laws that ultimately determine the macroscopic level on which irreversibility is such a familiar feature. Until now, many scientists have taken the view that the irreversibility observed in nature is merely illusory, and that the truth resides in the reversible dynamical laws which govern behaviour on a microscopic level (and of course is not directly observable). This state of affairs has come to prevail owing to an understandable but undue emphasis on simple, two-body problems in physics, which favour us sufficiently to provide exactly solvable models involving periodic motion. Once one leaves the domain of such simple models, and embarks on the study of systems with ever greater complexity, the former soon cease to provide an adequate paradigm in terms of which the latter can be understood. In short, complexity cannot be reduced to simplicity in any non-trivial manner: rather, irreversibility is seen to emerge as a necessary consequence of the instabilities present in more complicated systems.

A criticism that has not infrequently been aimed at the work of the Brussels School, however, is that, even if it is granted from a philosophical standpoint that irreversibility should be included in the fundamental level of description, there do not appear to be any new and experimentally verifiable consequences to have emerged from the approach. If this were the case, the proclaimed 'revolution' in scientific thought arising from the advent of an adequate role for irreversibility and time within the basic conceptual apparatus of theoretical physics would hardly be valid. I believe that there is at present some validity in this criticism, but hope it will be rectified by consideration of problems of concrete physical interest.

Although today the evidence indicates that the 'ultimate' description of matter must be framed in quantum-mechanical terms²⁷, quantum theory is not without its own difficulties: among these one may mention the non-exponential decay predicted for unstable particles, some general difficulties in the theoretical formulation of problems involving lifetimes²⁸, as well as the renormalization programme associated with the divergences of quantum field theory (even though this programme works well for many purposes)²⁹. It also appears that for problems necessitating the use of multi-photon detection apparatus for very short time-scales, a consistent theory of measurement is still required upon which to build a reliable interpretation of experimental observations³⁰. The approach I have described here offers a renewed possibility of clarifying some of these problems.

Modern cosmology uses thermodynamics in discussing the irreversible, evolutionary nature of the Universe: for example, one has only to consider the importance attached to the existence of the 3K microwave black-body radiation background. Another remarkable feature of contemporary cosmology, however, is its close relation with the reversible, non-evolutionary world of particle physics, previously embodied in the Standard Model of the 1970s, and now superseded by the Inflationary Universe description³¹. Fundamental progress in this area can be expected as the deeper connections between these two levels of description become clearer³².

From an epistemological viewpoint, the contributions of Prigogine's Brussels School are unquestionably of signal importance. The myth of a completely timeless, deterministic Universe is henceforth replaced by a world in which static affairs are enlarged to embrace the probabilistic kinetics of process; in which reversibility and irreversibility are accorded equal objectivity; and in which the notions of 'being' and 'becoming' are unified within a single conceptual framework¹⁹.

I thank Professor C. George for his remarks on the manuscript,

and for many helpful discussions. Thanks are also due to Professors I. Prigogine, B. Misra, J. Jeener, H. Swinney, and Drs F. Mayné, M. de Haan, V. Škarka, O. Némé, and A. Orłowska for stimulating conversations and enlightenment on various aspects of the work discussed here. Much of this article was

written at the Center for Statistical Mechanics at the University of Austin, Texas, USA (September 1986). I am also grateful to the Fondation Wiener-Anspach of l'Université Libre de Bruxelles for providing financial assistance.

1. von Helmholtz, H. *On the Interactions of the Natural Forces* (1854), reprinted in *Popular Scientific Lectures* (ed. Kline, M.) (Dover, New York 1961).
2. Eddington, A. S. *The Nature of the Physical World, Gifford Lectures 1927* (Cambridge University Press, 1928).
3. Onsager, L. *Phys. Rev.* **37**, 405-426 (1931); *ibid* **38**, 2265-2279 (1931). See also de Groot, S. R. & Mazur, P. *Non-equilibrium Thermodynamics* (North-Holland, Amsterdam, 1962).
4. Prigogine, I. *Etude Thermodynamique des Phénomènes Irréversibles* (Desoer, Liège, 1947).
5. Glansdorff, P. & Prigogine, I. *Thermodynamic Theory of Structure, Stability and Fluctuations* (Wiley-Interscience, New York, 1971).
6. Nicolis, G. & Prigogine, I. *Self-organisation in Non-equilibrium Systems* (Wiley-Interscience, New York, 1977).
7. Noyes, R. M. & Field, R. J. *A. Rev. Phys. Chem.* **25**, 95-119 (1974).
8. Taken from Roux, J.-C., Simoyi, R. H. & Swinney, H. L. *Physica D* **8**, 257-266 (1983), with permission.
9. Onsager, A. & Caplan, R. *A. Rev. Biophys. Bioeng.* **5**, 449-476 (1976).
10. Balescu, R. *Equilibrium and Non-equilibrium Statistical Mechanics* (Wiley-Interscience, New York, 1975).
11. Boltzmann, L. *Sitzungsber. K. Akad. Wiss. Wien* **66**, 275-370 (1872).
12. Tolman, R. C. *The Principles of Statistical Mechanics* (Dover, New York, 1979).
13. Prigogine, I. & Stengers, I. *Order Out of Chaos* (Heinemann, London 1984).
14. Poincaré, H. *C.R. Hebd. Séances Acad. Sci.* **108**, 550-553 (1889).
15. Arnold, V. I. & Avez, A. *Ergodic Problems of Classical Mechanics* (Benjamin, New York, 1968).
16. Misra, B. *Proc. natn. Acad. Sci. USA* **75**, 1627-1631 (1978).
17. Misra, B., Prigogine, I. & Courbage, M. *Proc. natn. Acad. Sci. USA* **76**, 3607-3611 (1979).
18. Bohr, N. *Atti Congr. Intern. Fis. Como* Vol. 2 (1927).
19. Prigogine, I. *From Being to Becoming* (Freeman, San Francisco, 1980).
20. Misra, B., Prigogine, I. & Courbage, M. *Proc. natn. Acad. Sci. USA* **76**, 4768-4772 (1979).
21. von Neumann, J. *Mathematical Foundations of Quantum Mechanics* (Princeton University Press, N.J., 1955).
22. Bohr, N. in *Quantum Theory and Measurement* (eds Wheeler, J. A. & Zurek, W. H.) (Princeton University Press, N.J., 1983).
23. Prigogine, I., George, C., Henin, F. & Rosenfeld, L. *Chem. Scr.* **4**, 5-32 (1973).
24. George, C., Prigogine, I. & Rosenfeld, L. *K. Dan. Vidensk. Selsk. Mat.-fys. Meddel.* **38**, 1-44 (1972).
25. Coveney, P. V. & George, C. 1988 (preprint).
26. Coveney, P. V. & George, C. *Physica* **141A**, 403-426 (1987); Coveney, P. V. *Physica* **143A**, 123-146 (1987); Coveney, P. V. *Physica* **143A**, 507-534 (1987).
27. d'Espagnat, B. *Une Incertaine Réalité* (Gauthier-Villars, Paris, 1985).
28. Heitler, W. *Quantum Theory of Radiation* (Oxford University Press, UK, 1954).
29. Dirac, P. A. M. *Directions in Physics* (Wiley-Interscience, New York, 1978).
30. Jeener, J. & Henin, F. *Phys. Rev. A* **34**, 4897-4928; Henin, F. & Jeener, J. *J. Stat. Phys.* **48**, 1321-1342 (1987).
31. Günzig, E. & Nardonne, P. *Fundamentals of Cosmic Phys.* **11**, 311-443 (1987).
32. Günzig, E., Géhéniau, J. & Prigogine, I. *Nature* **330**, 621-624 (1987).
33. Penrose, O. *Rep. Prog. Phys.* **42**, 1937-2006 (1979).
34. Škarka, V. & Coveney, P. V. *J. Phys. A* (in the press).
35. Coveney, P. V. *Adv. Phys.* (submitted).

ARTICLES

Chemical and isotopic evidence of thermochemical sulphate reduction by light hydrocarbon gases in deep carbonate reservoirs

H. Roy Krouse*, Christian A. Viau†, Leslie S. Eliuk†, Akira Ueda‡ & Stan Halas§

* Department of Physics, The University of Calgary, Calgary, Alberta T2N 1N4, Canada

† Shell Canada Limited, P.O. Box 2506, Calgary, Alberta T2P 3S6, Canada

‡ Mitsubishi Metal Corporation, Central Research Institute, 1-297 Kitabukuro, Omiya, Saitama 330, Japan

§ Marie Curie-Sklodowska University, Institute of Physics, Lublin, Poland

Large H₂S accumulations in deep carbonate reservoirs have been attributed to thermochemical sulphate reduction. Although organic matter has been invoked as the reducing agent, individual reactants in specific environments have hitherto not been identified. Carbon isotope data reveal that light hydrocarbon gases have been effective reducing agents in some Palaeozoic foothill reservoirs of western Canada.

GAS and oil fields contain H₂S as free gas or dissolved in formation waters. Concentrations vary from being just detectable to more than 90% of the gas phase¹⁻⁴. The high concentrations usually occur in deep carbonate reservoirs and represent an important resource for sulphur production worldwide⁵. It is generally accepted that such large H₂S accumulations arise through thermochemical sulphate reduction⁶⁻⁸ (TSR). Organic matter has been invoked as the reducing agent in general terms either directly⁹, or in a sequence involving reaction of H₂S with sulphate^{6,7}. It has been demonstrated in laboratory experiments that aqueous sulphate is reduced by a variety of organic compounds at temperatures above 250 °C (refs 10 and 11). On the basis of thermodynamic calculations, Anisimov⁴ concluded that significant accumulations of H₂S from the reduction of CaSO₄ by hydrocarbons in natural environments requires temperatures >150 °C. The crucial question is whether abiological sulphate reduction can occur at temperatures as low as 100 °C, that is, just above the limit of microbiological reduction. In a critical review, Trudinger *et al.*¹² concluded that abiological reduction below 200 °C had not been unequivocally demonstrated although they did not dismiss its possible geochemical sig-

nificance. But in laboratory experiments, Orr¹³ found that when H₂S was initially present, sulphate reduction occurred at lower temperatures. We have encountered subsurface environments where chemical and carbon isotope data can be used to argue effectively that TSR has occurred at temperatures <200 °C with light hydrocarbon gases serving as reducing agents. The data are from the Burnt Timber and Limestone Fields¹⁴ in the Upper Devonian Crossfield Member (Wabamun Group) and the Leduc Formation, as well as the Mississippian Turner Valley Formation, in the deep Foothills region (3,500-4,200 m) of south-western Alberta, Canada (Figs 1 and 2). These fields are thrust faulted anticlinal structures.

Chemical evidence of TSR

If sulphate is reduced by light hydrocarbon gases, then an extent of reaction parameter (ERP) can be defined by the molar ratio $[\text{CO}_2]/[(\text{CO}_2 + \sum_{n=1}^4 n\text{C}_n\text{H}_{2n+2})]$, which signifies the fraction of light hydrocarbons that have been oxidized. The multiplier factor, n , in the summation is required because n molecules of CO₂ are generated for each molecule of C_{*n*}H_{2*n*+2} consumed. This parameter does not account for CO₂ unrelated to oxidation

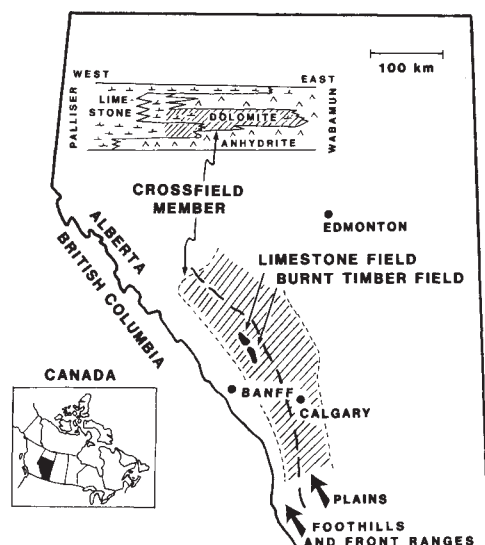


Fig. 1 Location of the Limestone and Burnt Timber Fields within the foothills of the Rocky Mountains in Alberta with the generalized stratigraphy of the Wabamun/Palliser Groups.

of hydrocarbons, nor for CO_2 removal from the system. It is seen that the mole per cent of H_2S increases with increasing ERP values (Fig. 3). For all samples, the molar concentrations of C_2H_6 , C_3H_8 and C_4H_{10} are <4%, 1% and 0.5%, respectively. Hence H_2S is associated with 'dry' gases. The two domains in Fig. 3 must reflect chemical differences in the two fields. This could be due to more CO_2 being present in the Limestone Field before the postulated TSR and/or more efficient CO_2 removal in the Burnt Timber Field. Regardless of these chemical differences the trends in Fig. 3 are consistent with the oxidation of $\text{C}_n\text{H}_{2n+2}$ to CO_2 with co-generation of H_2S .

Isotopic evidence of TSR

The $\delta^{34}\text{S}$ (CDT) and $\delta^{18}\text{O}$ (SMOW) values for anhydrites from the Crossfield Member at Burnt Timber and Limestone Fields ranged from +22.7 to +33.0‰ and from +12.0 to +19.2‰ respectively, with average values of +27.5 and +15.0‰. These are consistent with published data for late Devonian marine evaporites¹⁵. The H_2S gas was found to be more uniform in its $\delta^{34}\text{S}$ value (Fig. 3) and, on average, about 7‰ lighter than the evaporites. Several anhydrite samples, however, were obtained from a given cored formation at different depths in contrast to only one gas sample. The difference in $\delta^{34}\text{S}$ values between H_2S and anhydrite is smaller than the 14–22‰ measured during chemical SO_4^{2-} reduction in the laboratory from below room temperature up to 150 °C with powerful reducing agents not encountered naturally^{11,16,17}. The fractionation was attributed to the kinetic isotope effect during initial S—O bond rupture¹⁶. This implies that a reaction step with lower isotopic selectivity was rate-controlling in the Foothills reservoirs. If sulphate reduction occurred immediately on dissolution of anhydrite (Fig. 4b), the overall fractionation should be small because dissolution is essentially a surface reaction. The fractionation of about 7‰ suggests that reduction was not immediate, that is, the concentration of SO_4^{2-} was sufficient to permit some isotopic selectivity but the kinetic isotope effect for S—O bond rupture was not fully expressed.

Elemental sulphur was found to be slightly more enriched in ^{34}S than the H_2S . This could reflect fractionation during oxidation of H_2S , or kinetic isotope effects during processes such as dissociation of polysulphides. Another possibility is that the S^0 was generated from the reaction of H_2S with sulphate^{6,7,13,18}, in which case the S^0 should display a $\delta^{34}\text{S}$ value intermediate to the H_2S and the sulphate¹⁸.

The most compelling evidence for TSR by light hydrocarbon

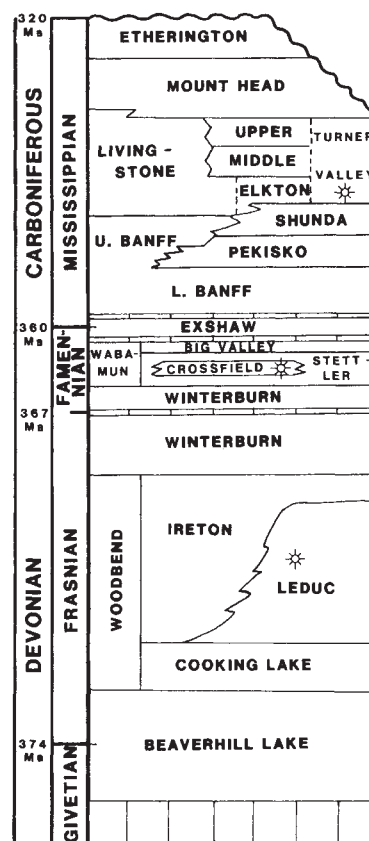


Fig. 2 Middle and Upper Devonian and Lower Carboniferous stratigraphy in southern Alberta. Ages are from Harland *et al.*³⁴. The sources of gases used in this study (Crossfield, Leduc and Turner Valley units) are indicated by the symbol *.

gases is in their isotopic composition. In particular, the $\delta^{13}\text{C}$ (PDB) values for ethane and propane are about 10‰ higher than previously found for Devonian gases¹⁹ (Fig. 5). The $\delta^{13}\text{C}$ values for methane are comparable to the highest found elsewhere in the Devonian. For the Crossfield Member, there are distinct trends of C_2H_6 and C_3H_8 becoming more enriched in ^{13}C with increasing ERP (Fig. 6). A similar, but lower, sloping trend is evident with CH_4 . There is a weaker trend of CO_2 becoming more depleted in ^{13}C with increasing ERP. These trends are consistent with partial oxidation of the light hydrocarbon gases, whereby the ^{12}C -bearing species reacted preferentially. Hence the unreacted hydrocarbon gases became more enriched in ^{13}C and ^{13}C -depleted product CO_2 was added to the original CO_2 . The $\delta^{13}\text{C}$ values for CO_2 are not as negative as predicted by a simple mixing model because of carbon isotope exchange with dissolved carbonate species which in turn are partly generated through dissolution of marine carbonates (Fig. 4e).

It was also found that data from gases in other geological units (Mississippian Turner Valley and Devonian Leduc Formations) of the same wells are consistent with the trends for the Crossfield Member gases (Fig. 6). This implies that the TSR proceeded in a system that remained closed with respect to hydrocarbon gases.

There are two isotopic populations of carbonates (Fig. 6). The higher $\delta^{13}\text{C}$ values of the dolomites are those associated with precipitation from fluids with a near-marine signature. The more negative $\delta^{13}\text{C}$ values of the secondary calcites must reflect CO_2 generated by the oxidation of hydrocarbon gases. The lowest $\delta^{13}\text{C}$ value for secondary calcite of -29‰ does not rule out bulk organic matter as a source of carbon; but considering the evidence for dissolution of marine carbonate (Fig. 4e), the $\delta^{13}\text{C}$ value of -29‰ probably does not represent oxidized

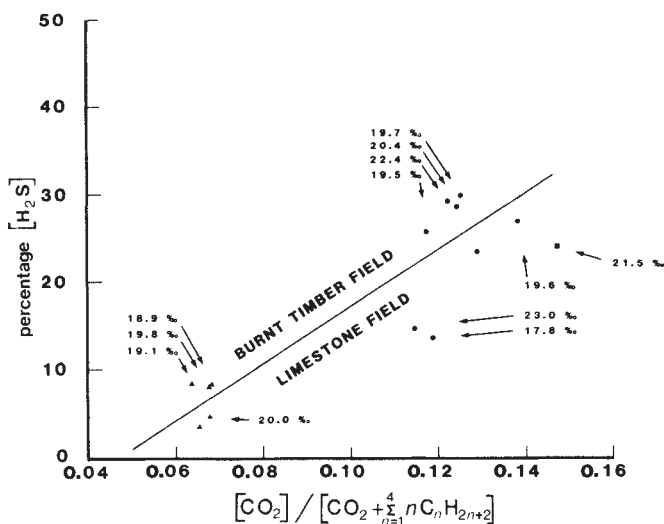


Fig. 3 H_2S concentration in produced gas from Limestone and Burnt Timber Fields plotted against hydrocarbon composition of the produced gas in terms of the extent of thermochemical sulphate reduction parameter defined as $[\text{CO}_2] / [\text{CO}_2 + \sum_{n=1}^4 n \text{C}_n \text{H}_{2n+2}]$. $\delta^{34}\text{S}$ values of H_2S (for most samples) are shown by arrows. Triangles, Mississippi Turner Valley Formation; circles, Devonian Crossfield Member; square, Devonian Leduc Formation. Isotope abundance in Figs 3, 5 and 6 are expressed on a δ -scale (del) defined as

$$\delta_2 = 10^3 \left\{ \frac{X_2/X_1}{S_2/S_1} - 1 \right\} \%$$

where 1 and 2 refer to the light and heavy isotopes, and X and S refer to the unknown and standard samples, respectively. Laboratories throughout the world have agreed on the standard to be used as the zero for the δ -scale of a given pair of isotopes.

organic matter, but a mixture of CO_2 , as described above. Hence the carbon isotope data for calcite imply indirectly that light hydrocarbon gases with $\delta^{13}\text{C}$ lower than -29% served as reducing agents.

Temperature estimates

Estimation of the temperature of reaction is difficult because of the inability to measure it directly and the uncertainty as to the time that the bulk of the reaction occurred. But the available evidence strongly suggests that the reaction was abiological and occurred between 90°C and 175°C .

The high concentrations of H_2S (Fig. 3) are not characteristic of a microbial system. Where the H_2S production rate is high, bacterial survival requires its removal by precipitation of sulphide minerals or bacterial oxidation to S^0 and SO_4^{2-} as in a 'sulphuretum'²⁰.

In a survey of $\delta^{34}\text{S}$ values for H_2S in Devonian carbonates²¹, there was a marked transition at current reservoir temperatures near 80°C . Below this temperature, fewer positive $\delta^{34}\text{S}$ values were found which were attributed to possible biological participation. Above this temperature, the values were close to those of associated evaporites and attributed to thermochemical sulphate reduction. The current data are similar to those in the higher-temperature category.

Had bacteria participated, greater extremes in the $\delta^{13}\text{C}$ values of secondary carbonates would be expected, as found for example in the Texas-Louisiana salt domes²²⁻²⁴. Large and variable shifts in the $\delta^{13}\text{C}$ values of propane and ethane in gases with minimal H_2S have been attributed to bacterial conversion²⁵. The smaller shifts and lesser variability combined with high H_2S content of these data are more consistent with abiological reaction. The chemical composition of the light hydrocarbon gases (presence of ethane, propane and so forth) is characteristic of thermogenesis. Further, the $\delta^{13}\text{C}$ values of these gases, although

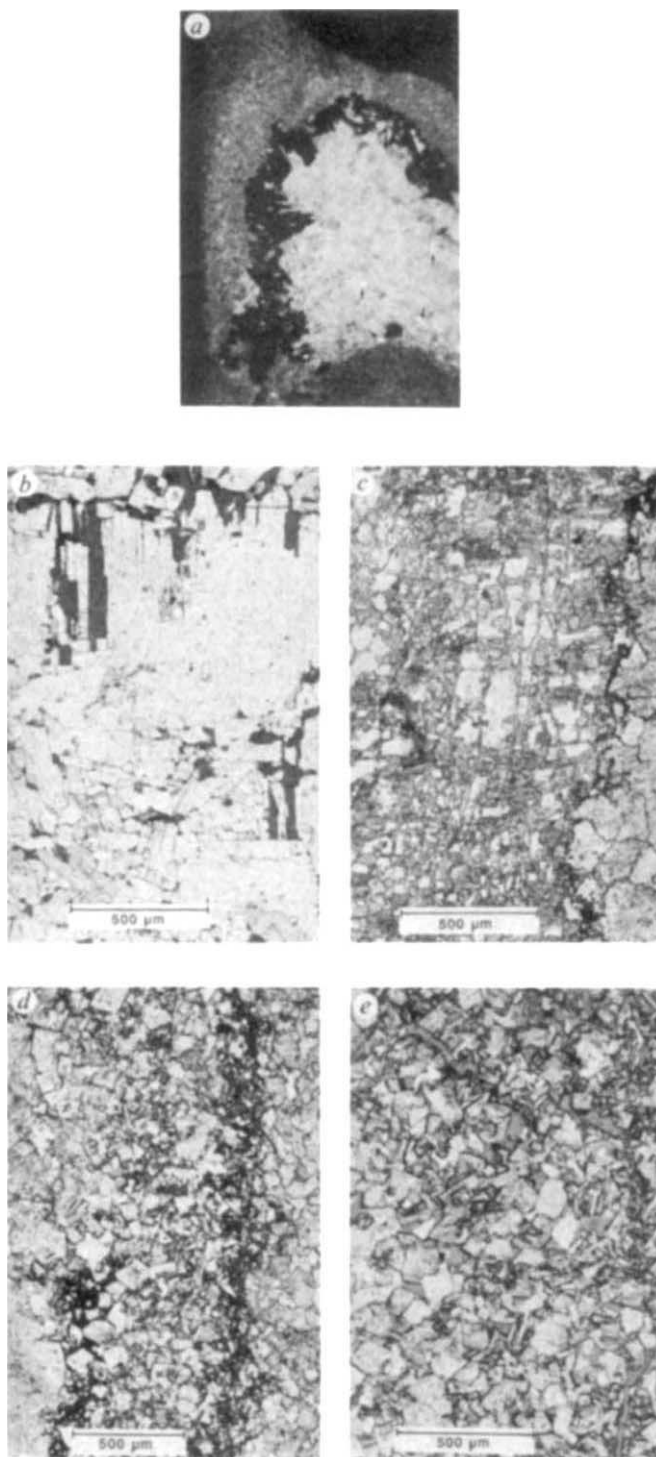


Fig. 4 *a*, Polished core slab showing typical anhydrite nodule (white area) partly replaced by calcite (dark-stained crystals) and elemental sulphur (bright areas shown by arrows) and surrounded by porous dolomite. Width of photograph is 1.5 cm. *b*, Thin-section photomicrograph showing crystallographically controlled dissolution of anhydrite crystals with partial collapse of resulting blades at the base. Bar scale: 0.5 mm. *c*, Calcite (dark-stained material) replacing anhydrite crystal. The isolated anhydrite (light coloured) areas within the calcite are all in optical continuity. Note the orthogonal nature of the replacement along former anhydrite cleavage. Bar scale: 0.5 mm. *d*, Porous dolomite rim around 'nodule' composed of calcite replacing anhydrite. Area on the right is composed of non-porous dolomite. Bar scale: 0.5 mm. *e*, Area of porous dolomite showing intercrystalline porosity associated with the unusual crystal-shaped dolomite due to dissolution. Bar scale: 0.25 mm.

modified, are those associated with thermogenesis²⁶. The temperature of formation of the hydrocarbon gases may not be indicative of that of the sulphate reduction reaction; but these gases were trapped in a structure formed during or after the Laramide orogeny²⁷ (late Jurassic to Palaeocene in age). It is unlikely that the temperature in the reservoirs was ever lower than the current reservoir temperatures of 89–95 °C. Although some bacteria have been reported as surviving at such temperatures¹², it is unlikely that sulphate-reducing bacteria can function in this range¹². Thus we conclude that the lower-temperature limit for the reaction was ~90 °C and sulphate-reducing bacteria did not participate.

To establish the range of reaction temperatures, fluid inclusion homogenization temperatures were determined in secondary calcites. Homogenization temperatures range from 36 °C–175 °C, but H₂S, CO₂ and possibly CH₄ occur in these inclusions. Homogenization temperatures for such complex mixtures depends strongly on composition and are not representative of temperatures of formation. Because these systems have not been studied, the range 36–175 °C cannot be properly interpreted, but some fluid inclusions were found to contain evidence for trapping of S⁰ as an immiscible liquid. This indicates temperatures ≥90 °C (ref. 28), depending on the pressure and composition of gases.

Indices of maturity (Ro and LOM) and considerations of burial history suggest a temperature range of 140 °C–175 °C; therefore, 175 °C would seem to be the upper temperature limit. Although, Trudinger *et al.*¹² felt that there was not compelling evidence for laboratory TSR below 200 °C, Orr¹³ reported that the presence for H₂S promoted the reaction. He found the half-life of SO₄²⁻ with an initial H₂S pressure of 200 p.s.i. (≈1,379 kPa) to be about 1,800 h at 175 °C. He further speculated that SO₄²⁻ reduction by the combination of H₂S and organic matter could occur at geologically significant rates down to 100 °C.

Petrographic modifications by TSR

The Crossfield Member consists of a porous dolomite with intercrystalline pores regularly distributed or showing an irregular patchy distribution in an otherwise tight dolomite, giving the core a mottled appearance. Secondary anhydrite nodules are commonly associated with porosity, calcite and elemental sulphur^{14,29} and are, in places, surrounded by porous dolomite (Fig. 4a). Figure 4b shows partly dissolved, bladed, anhydrite crystals, some of which have collapsed. The dissolution conforms to growth orientation, an indication of surface-controlled rather than transport-controlled dissolution³⁰. Figure 4c shows calcite replacing anhydrite along cleavage planes. The relict anhydrite is optically continuous confirming that it was a single entity before replacement. Figure 4d shows the edge of an anhydrite nodule where calcite has replaced anhydrite. The porous dolomite surrounding the nodule is separated from the

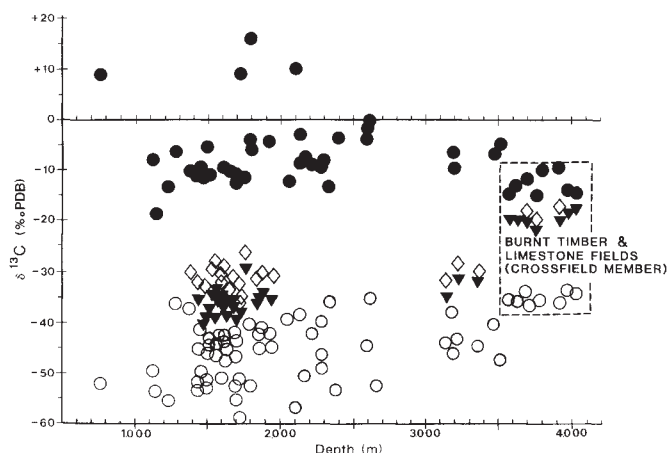


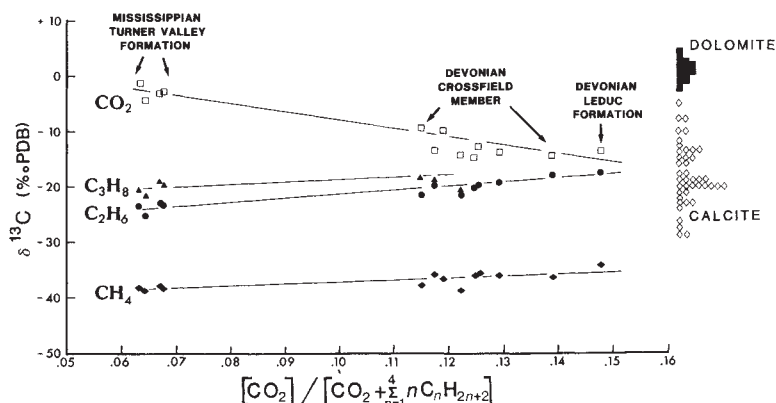
Fig. 5 Values of $\delta^{13}\text{C}$ for methane (open circles), ethane (filled triangles), propane (diamonds) and carbon dioxide (filled circles) gases produced from ten different Devonian stratigraphic units in the Alberta Basin are plotted against depth. Values for samples from the Crossfield Member at Limestone and Burnt Timber Fields are in the box outlined on the right-hand side (depth greater than 3,500 m). Ethane and propane values are anomalously heavy by ~10‰, as compared with other Devonian gases reflecting their participation in TSR. Methane values are in the upper range of Devonian values, whereas carbon dioxide values are among the most negative values encountered in the Devonian. The data at depths <3500 m are from Krouse¹⁹.

calcite and tight dolomite by black carbonaceous residues. The porous dolomite is composed of fragments of subunits of original dolomite rhombs that can only be explained in terms of dissolution (Fig. 4e). The dissolution process is interpreted to be related to the TSR reaction for the following reasons: (1) the nature of dolomite dissolution is very similar to that of anhydrite in that it conforms to growth orientation and hence is a surface-controlled process, (2) the porous dolomite is found in close association with sulphate reduction reactants and products, (3) there is a positive correlation between the porosity-thickness in different Crossfield Member reservoirs and the concentration of H₂S present²⁹. In addition, reprecipitation of dolomite is observed. Consequently, creation, occlusion and redistribution of porosity is not only related to the dissolution of anhydrite and precipitation of elemental sulphur and calcite, but redistribution of porosity appears to be related to dissolution-reprecipitation of the dolomite reservoir itself.

Conclusions

Petrographic, gas composition and isotopic data are all consistent with thermochemical reduction of sulphate by light hydrocarbon gases. The large H₂S accumulations in the Crossfield Member reflect the extent of this reaction. Indeed,

Fig. 6 Values of $\delta^{13}\text{C}$ for methane, propane and carbon dioxide gases from the Mississippian Turner Valley Formation, Devonian Crossfield Member and Devonian Leduc Formation plotted against the extent of (TSR) reaction parameter. The diagram shows the inverse relation between $\delta^{13}\text{C}$ of CO₂ and that of other gases, particularly ethane and propane. This is interpreted, along with the anomalous composition of these gases with respect to other Devonian gases (shown in Fig. 5), as evidence that natural gases presently in the reservoirs have served as reactants in the thermochemical reduction of sulphate. Values of $\delta^{13}\text{C}$ for reservoir dolomite and calcite are shown on the right-hand side. The negative values for calcite reflect substantial incorporation of organic carbon from the isotopically light gas, whereas the more positive values reflect various mixtures of organic carbon and carbon (with near 0‰ composition) released by the dissolution of dolomite.



the modifications of reaction components during the extensive H_2S production provided the key for better characterization of the TSR process. Furthermore, it is commonly observed that high H_2S concentrations are associated with methane-dominated gases. The greater shifts in $\delta^{13}\text{C}$ for ethane and propane in comparison to methane, in this study, suggest that the former were oxidized more extensively. Whereas light hydrocarbon gases are identified as the dominant reducing agents in the Crossfield Member, sulphate reduction by organic matter of higher molecular mass may prevail in other environments.

Abiological sulphate reduction apparently occurred between 90 °C and 175 °C. Because the reactants have not been exhausted, TSR could conceivably be occurring near 100 °C, the current reservoir temperature. In any case, this study suggests that TSR by light hydrocarbon gases occurred at temperatures <200 °C.

The porous network in the Crossfield Member has been significantly affected by the diagenetic process leading to accumulation of large quantities of H_2S . Dissolution of anhydrite and precipitation of elemental sulphur and calcite have created and to a lesser extent occluded porosity. The dolomite reservoir itself cannot be considered as a passive part of the system; dolomite

dissolution and reprecipitation have created and mostly redistributed porosity.

The findings in the Crossfield Member have clear implications for the use of carbon isotopes in gas exploration. Much effort has been devoted to interpreting plots of $\delta^{13}\text{C}$ versus chemical composition²⁶, $\delta^{13}\text{C}$ versus indices of maturity (vitrinite reflectance³¹, Ro; level of organic metamorphism³², LOM), and $\delta^{13}\text{C}$ versus δD (ref. 33), to delineate continental and marine organic sources and identify source rocks. Gases that have participated in TSR are likely to have modified chemical compositions and higher $\delta^{13}\text{C}$ values. Hence the data will not fit the established trends associated with cracking reactions. For example, on the basis of $\delta^{13}\text{C}$ values, a source rock with higher maturity of organic matter would be predicted than is actually so.

This work was supported by grants from Shell Canada Limited, Shell Research International Company (SIRM) and the Natural Sciences and Engineering Research Council of Canada. We thank Mary Given, Charlie Bruce and Felix Frey (Shell Canada) for discussions, critical review of this manuscript and support for the project; also Kathy Aulstead for fluid inclusion analyses. Comments by Max Coleman helped to improve the manuscript.

Received 6 October 1987; accepted 2 March 1988.

- Hitchon, B. J. *Can. petrol. Technol.* **2**, 100–116 (1963).
- Anisimov, L. A. *Int. geol. Rev.* **13**, 191–197 (1971).
- Hyne, J. B., Lau, C. S. C., Dowling, N. I., Davis, P. & Lesage, K. *Alberta Sulphur Research Ltd Report XXII No. 1*, 25–39 (1985).
- Anisimov, L. A. *Geochem. int.* **15**, 63–71 (1978).
- Harben, P. *Mining Mag.* **155**, 374–379 (1986).
- Orr, W. L. *Am. Assoc. petrol. geol. Bull.* **50**, 2295–2318 (1974).
- Orr, W. L. *Advances in Organic Geochemistry 1975, Madrid, Enadisma* (eds Campos, R. & Goni, J.) 571–597 (1977).
- Krouse, H. R. *J. geochem. Explor.* **7**, 189–211 (1977).
- Bely, V. M. & Vinogradov, V. I. *Geologiya Nefti Gaza N2*, 37–41 (1972).
- Toland, W. G. *J. Amer. chem. Soc.* **82**, 1911–1916 (1960).
- Kiyosu, Y. *Chem. Geol.* **30**, 47–56 (1980).
- Trudinger, P. A., Chambers, L. A. & Smith, J. W. *Can. J. Earth Sci.* **22**, 1910–1918 (1985).
- Orr, W. L. *Geol. Soc. Am. Abstr. Progr.* **14**, 580 (1982).
- Eliuk, L. S. & Hunter, D. F. in *13th Can. Soc. Petrol. Geol. Core Conf.* (eds F. Krause and G. Burrows) 39–62 (1987).
- Claypool, G. E., Holser, W. T., Kaplan, I. R., Sakai, H. & Zak, I. *Chem. Geol.* **28**, 199–260 (1980).
- Harrison, A. G. & Thode, H. G. *Trans. Faraday Soc.* **53**, 1648–1651 (1957).
- Grinenko, V. A., Grinenko, L. N. & Zagryazhskaya, G. D. *Geokhimiya* **4**, 484–491 (1969).
- Husain, S. A. & Krouse, H. R. New Zealand Department Scientific and Industrial Research Bull. 220, 207–210 (1978).
- Krouse, H. R. Final Report Alberta-Canada Energy Resources research fund, agreement U-30, 145 pp. (1983).
- Baas Beeking, L. G. M. *Ann. Bot.* **39**, 613–650 (1925).
- Krouse, H. R. in *Proc. 10th World Petrol. Congress* Vol. 10, 85–92 (1980).
- Thode, H. G., Wanless, R. K. & Wallouch, R. *Geochim. cosmochim. Acta* **5**, 286–298 (1954).
- Feely, H. W. & Kulp, J. L. *Am. Assoc. petrol. geol. Bull.* **41**, 1802–1853 (1957).
- Davis, J. B. & Kirkland, D. W. *Econ. Geol.* **65**, 107–121 (1970).
- James, A. T. & Burns, B. J. *Am. Assoc. petrol. geol. Bull.* **68**, 957–960 (1984).
- Stahl, W. *Chem. Geol.* **20**, 121–149 (1977).
- Bally, A. W., Gordy, P. L. & Stewart, G. A. *Bull. Can. Petrol. Geol.* **14**, 337–381 (1966).
- Woll, W. *Erdöl Kohle Erdgas Petrochem.* **99**, 297–300 (1983).
- Eliuk, L. S. 1984 *Can. Soc. Petrol. Geol. Core Conf.* 245–289 (1984).
- Berner, R. A. *Am. J. Sci.* **278**, 1235–1252 (1978).
- Stahl, W. & Carey, B. C. *Chem. Geol.* **16**, 257–267 (1975).
- James, A. T. *Am. Assoc. Petrol. Geol. Bull.* **67**, 1176–1191 (1983).
- Schoell, V. M. *Am. Assoc. Petrol. Geol. Bull.* **67**, 2225–2238 (1983).
- Harland, W. B. et al. *A Geologic Time Scale: Cambridge Earth Science Series* (Cambridge Univ. Press, 1982).

Morphological structures produced by mixing in chaotic flows

J. M. Ottino*, C. W. Leong*, H. Rising† & P. D. Swanson‡

* Department of Chemical Engineering, University of Massachusetts, Amherst, Massachusetts 01003, USA

† Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts 01003, USA

‡ Department of Polymer Science and Engineering, University of Massachusetts, Amherst, Massachusetts 01003, USA

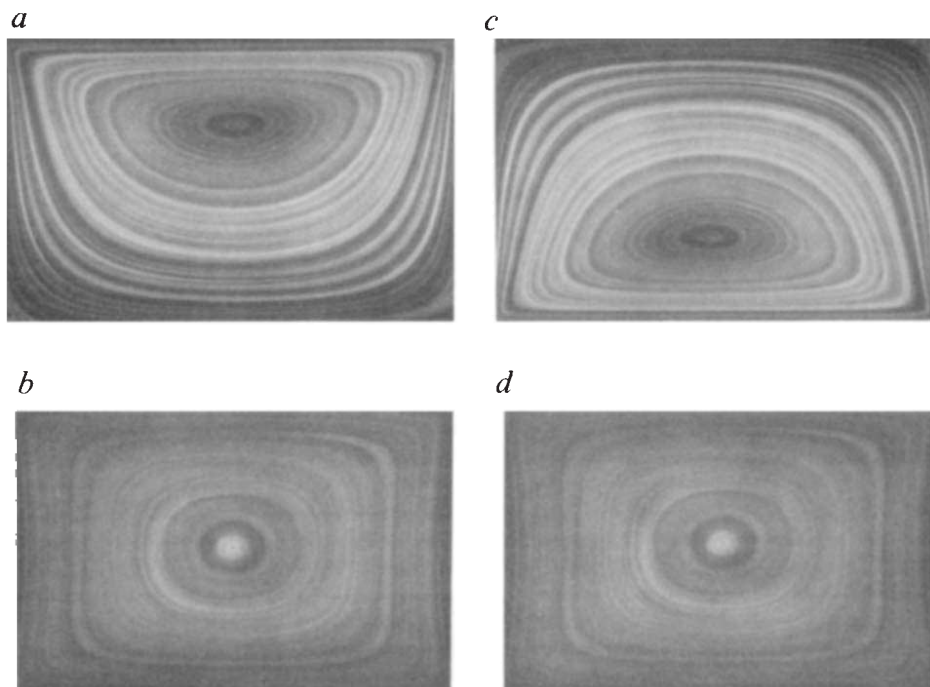
Fluid mixing is poorly understood. However, some of the main concepts can be captured by both computational and experimental studies of two-dimensional flows. The identification of coherent structures and how they change with variations of the governing parameters gives insight into the behaviour of chaotic systems and can serve as a model for a variety of problems found in nature.

A STUDY of chaotic mixing of fluids in two-dimensional time-periodic flows is important for two reasons. First, it serves as a useful model for the understanding of a variety of mixing processes. Second, it provides a visual analogue for chaos in area-preserving maps, one of the oldest and most established systems to be analysed in the context of dynamical systems^{1,2}.

Fluid mixing consists of the stretching and distribution of fluid elements throughout space; alternatively, it can be viewed as the creation of interfacial area and reduction of length scales between initially segregated fluids. In the most general case,

mass diffusion and possibly chemical reactions occur concurrently with the fluid mechanical processes of stretching and folding. Fluid mixing serves also as a physical analogue for many important concepts in dynamical systems theory and chaos. Indeed if we restrict our attention to mixing occurring in closed, time-periodic, two-dimensional flows (no material enters or leaves the system), the stretching and folding of fluid elements corresponds to what in dynamical systems is called a Smale horseshoe^{2–4} and which is also one of the oldest concepts in mixing⁵.

Fig. 1 Instantaneous picture of the streamlines corresponding to the prescription of equation (5); *a*, corresponds to $v_{\text{top}} = U$ and $v_{\text{bot}} = 0$; *b*, to $v_{\text{top}} = -v_{\text{bot}}$, *c*, to $v_{\text{top}} = 0$ and $v_{\text{bot}} = -U$, and *d*, with $v_{\text{top}} = -v_{\text{bot}}$. The photographs correspond to the case $U = 2.69 \text{ cm s}^{-1}$. The dimensions of the cavity, constant in all the experiments, are $W = 10.3 \text{ cm}$ and $H = 6.2 \text{ cm}$. The fluid is glycerine at 25°C .



A horseshoe is one of the possible signatures of chaos. More generally, a system can be labelled 'chaotic' by means of various computational, analytical and experimental diagnostics. Although the application of the diagnostics depends largely on the nature and complexity of the system studied, most 'chaotic' systems have been labelled as such by computational means.

In the context of the models described here a system can be labelled as chaotic if it satisfies any of the following, roughly equivalent, criteria: (1) the flow forms horseshoes, (2) the flow contains transverse homoclinic or heteroclinic intersections, and (3) the flow has a positive Liapunov exponent. All definitions are described in detail in ref. 9. A somewhat imprecise definition of (3) is that fluid particles, no matter how close initially, separate exponentially fast.

Here we shall examine some of the structures produced by experimental studies in model time-periodic flows, which according to the definition (1) can be labelled as chaotic⁴. Such systems provide useful insight into the behaviour of chaotic systems. Indeed, fluid mixing provides one of the few instances where it is possible to identify chaotic processes in terms of directly observable physical phenomena. We observe coherent regions (holes) which translate and deform with the flow but where there is little mixing. These regions are surrounded by 'chaotic' regions where mixing is effective. The behaviour of these structures in simple flows might help to understand various mixing processes occurring in nature and technology.

Theory

Consider the eulerian velocity field $\mathbf{v}(\mathbf{x}, t)$. The trajectory of a fluid particle initially located at $\mathbf{x} = \mathbf{x}_0$ corresponds to the solution of the dynamical system

$$d\mathbf{x}/dt = \mathbf{v}(\mathbf{x}, t) \quad (1)$$

with the initial condition $\mathbf{x} = \mathbf{x}_0$. The solution of equation (1) for all $\mathbf{x}_0$ is called the flow⁶

$$\mathbf{x} = \phi_t(\mathbf{x}_0). \quad (2)$$

In continuum mechanics the transformation (2) is called the motion⁷. In two dimensions $\mathbf{v} = (v_x, v_y)$, $\mathbf{x} = (x, y)$, and the flow $\mathbf{x} = \phi_t(\mathbf{x}_0)$ corresponds to an area-preserving map. The system (1) has a hamiltonian structure⁸

$$dx/dt = \partial\psi/\partial y \quad dy/dt = -\partial\psi/\partial x \quad (3)$$

where ψ is the streamfunction (for hamiltonian systems, see refs 9–11). It is possible to assert that if the velocity field is steady, that is, ψ is time-independent, the velocity field is integrable and the system cannot be chaotic. If, however, the velocity field, or equivalently ψ , is time-dependent, hamiltonian mechanics indicates that there is a good chance that the system will be chaotic, and indeed such systems appear to be common in practice^{3,4,8}. In the special case of a time-periodic velocity field where $\mathbf{v}(\mathbf{x}, t) = \mathbf{v}(\mathbf{x}, t + T)$, the flow given by equation (2) can be reduced to a mapping

$$\mathbf{x}_{n+1} = \phi_T(\mathbf{x}_n) \quad (4)$$

where we identify the position of $\mathbf{x}_0$ at time $t = T$ as $\mathbf{x}_1$, at time $t = 2T$ as $\mathbf{x}_2$, and so on.

Fluid mixing occurs in a closed system, and this fact alone has several important consequences such as Poincaré recurrence⁶. Some fluid particles return exactly to their original positions whereas others recur, that is, pass arbitrarily close to their initial locations. For the system described by equation (4), if a designated particle at $t = 0$ returns exactly after one period, that is at $t = T$, but not before then, the initial location corresponds to a periodic point of period one; if it returns after two periods, $t = 2T$, but not before then, the point is of period two, and so on. The superposition of stroboscopic pictures of the location of the positions of various initial conditions $\mathbf{x}_0$ taken at times nT , with $n = 0, 1, 2, \dots, N$, for large N , shows the evolution of a Poincaré section of the initial conditions $\mathbf{x}_0$. Such a picture represents the behaviour of the system over a long time for all initial conditions (even though the number of initial conditions is finite, if they are selected properly, they show the existence of structures which determine qualitatively the character of the flow for 'all' initial conditions).

As discussed below, the key to understanding the complex behaviour of chaotic flows resides in the character of periodic points. The periodic points in the Poincaré section can be classified as hyperbolic, parabolic or elliptic, according to the deformation of the fluid in the neighbourhood of a periodic point (the parabolic being a degenerate case). The elliptic points are surrounded by islands: invariant regions which translate and rotate conserving identity but do not exchange matter with the rest of the fluid; they represent an obstacle to efficient mixing. The hyperbolic points have associated stable and unstable manifolds; when they intersect transversely they form homoclinic

and/or heteroclinic orbits (hyperbolic sets⁹). The large-scale mixing depends critically upon the interaction of manifolds belonging to different periodic points³. Observationally, the stretching rate within islands is nearly linear; the stretching rate in chaotic regions can be exponential. An exponential stretching is desirable from the viewpoint of mixing because it implies the fastest generation of intermaterial area between fluids.

Mathematically, a system can be classified as chaotic if it is able to produce horseshoes or transverse homoclinic intersections. A horseshoe implies infinitely many periodic points and transverse homoclinic points; likewise a transverse homoclinic point implies infinitely many horseshoes. When the system is near-integrable^{10,11}, or in the neighbourhood of elliptic points, various celebrated theorems give information about the structure of the system, for example, the KAM (Kolmogorov-Arnold-Moser) theorem and Moser's twist theorem. The simplest problem corresponds to the case when the map is known explicitly. In this case, and to a lesser degree when the streamfunction (see equation (3)) is known explicitly or when the perturbation from integrability is small (near-integrable system), an array of analytical and computational techniques are possible such as the determination of homoclinic and heteroclinic intersections by means of the Melnikov technique⁹. However, the analysis of chaotic mixing has two important differences from conventional studies of chaotic systems: (1) in mixing we are interested in rate processes rather than asymptotic behaviour, that is, we are interested in phenomena which occur over relatively few periods, and (2) generally, we are interested in large perturbations from integrable systems rather than near-integrable behaviour because this is when the best fluid mixing occurs.

The theory is considerably less developed for these two cases and one of the objectives of the work described here is to put forward the idea that experiments can contribute to the theoretical development on this subject. In the field of dynamical systems, computer simulations have become a 'laboratory' for 'experimental mathematics'. The actual fluid system provides an alternative visual tool, less controllable perhaps, but with the advantage that much of the behaviour of the system can be directly seen, particularly in the regions of hyperbolic sets with the aid of a tracer injected into the fluid. Observations are limited to macroscopic objects produced by fluid mixing, and the question is how these objects behave under modification of a forcing parameter (T) taken as an arbitrary bifurcation parameter.

Two-dimensional time-periodic flows

Here, we illustrate some of the above concepts by conducting some laboratory experiments of two time-periodic two-dimensional systems: a cavity-flow and a journal-bearing flow.

The experiments described here were conducted in an improved computer-controlled version of the system described by Chien *et al.*⁴ (C.W.L. and J.M.O., manuscript in preparation). The system consists of two sets of roller-pairs connected by timing belts driven independently by reversible motors, and two neoprene bands that act as moving walls. The entire system is immersed in a plexiglass tank of depth ~ 25 cm and entirely filled with glycerine. The Reynolds number used in the experiments is the highest compatible with two competing effects: two-dimensionality (absence of inertia and secondary flows) and minimum diffusion of the dye (tracer) during the experiment. An order of magnitude calculation based on diffusional effects gives a Reynolds number ~ 1 . Both the Reynolds number (Re) and the Strouhal number (Sr) which are a measure of the inertial effects in the flow, are kept low enough in the experiments as to be of no importance. (In the computation of Re and Sr the characteristic velocity is taken as U and the characteristic length as H^2/W where H is the height and W the width of the rectangular region comprising the system.)

The cavity-flow apparatus is used to study two classes of time-periodic flows: (1) a discontinuous operation of the boundaries in a co-rotating sense, that is, jumping between Fig. 1*a, c*;

the top and bottom walls move for a time $T/2$ at a constant speed U in opposite directions; and (2), a periodic flow corresponding to wall motions of the form

$$v_{\text{top}} = U_{\text{top}} \sin^2(\pi t/T), \quad v_{\text{bot}} = -U_{\text{bot}} \sin^2(\pi t/T + \pi/2) \quad (5)$$

Thus, the system evolves smoothly in the direction of Fig. 1*a-d*. We regard the period T as the governing parameter or, equivalently, the wall displacements d_{top} , d_{bot} , during the time T

$$d_{\text{top}} = \int_0^T v_{\text{top}}(t) dt, \quad d_{\text{bot}} = \int_0^T v_{\text{bot}}(t) dt \quad (6)$$

Owing to its experimental versatility, the cavity flow allows us to investigate a variety of mixing conditions. However, it is not well suited for numerical studies because the streamfunction is not known analytically. On the other hand, the journal-bearing flow, which is the motion between two co-rotating or counter-rotating eccentric cylinders has an exact solution for creeping flow, can be simulated numerically and provides a bridge between experiment and theory. The main parameters of this system are two geometrical ratios, the ratio of radii of the inner and outer cylinders $R_{\text{in}}/R_{\text{out}}$, and the degree of eccentricity (distance between centres of the cylinders, δ), measured as δ/R_{out} . Under creeping flow conditions¹³, the streamline portrait is determined only by the ratio of angular velocities of the inner and outer cylinders, $\Omega_{\text{in}}/\Omega_{\text{out}}$. The numerical simulations we carried out are based on the solution corresponding to the creeping-flow case obtained by Wannier¹⁴. According to the operating conditions the flow can display one or two saddle points. We focus, exclusively, on a counter-rotating case which under steady state displays one saddle point. The parameter values are $R_{\text{in}}/R_{\text{out}} = 1/3$, $\delta/R_{\text{out}} = 0.3$ and $\Omega_{\text{in}}/\Omega_{\text{out}} = -2$.

The most important design feature of our apparatus compared with previous designs¹⁵ is that it allows an unobstructed view of the flow region because the outer cylinder is rotated from 'outside' (with a large bearing enclosing the outer cylinder). This is particularly important in the case under study because the period-1 hyperbolic point would not be visible in conventional designs. Observations are made from the bottom of the apparatus which is made of glass. The working fluid is glycerine floated on carbon tetrachloride.

The operation of the journal-bearing flow is similar to that of the cavity flow. The solution corresponding to the creeping-flow problem is given by $\nabla^4 \psi = 0$. As this problem is linear, the solution can be written as a linear combination of the forcings of the contributions corresponding to the inner and outer cylinders:

$$\psi = \psi_{\text{in}}(x, y)\Omega_{\text{in}} + \psi_{\text{out}}(x, y)\Omega_{\text{out}} \quad (7)$$

which suggests the adoption of a pseudo-steady-state viewpoint for the time-periodic case¹⁶,

$$\psi = \psi_{\text{in}}(x, y)\Omega_{\text{in}}(t) + \psi_{\text{out}}(x, y)\Omega_{\text{out}}(t) \quad (8)$$

The necessary conditions for the experiments to match the computer simulations are $Re \ll 1$ and $Sr \ll 1$. Combining the two conditions, we require that the timescale of the perturbation (T) be much larger than the timescale of the diffusion of momentum, that is, $T \gg L^2/\nu$ (ν is the kinematic viscosity). The computations are based on numerical solutions of equation (3) where $\Omega_{\text{in}}(t)$ and $\Omega_{\text{out}}(t)$ act as input to the system through equation (8). First, we notice that the streamfunction is independent of the actual speed of the boundary. This implies that, as long as the velocity histories do not overlap, that is, $\Omega_{\text{in}}(t) = 0$ whenever $\Omega_{\text{out}}(t) \neq 0$ and vice versa, any numerical results will be identical provided that the angular displacements per period θ_{in} , θ_{out}

$$\theta_{\text{in}} = \int_0^T \Omega_{\text{in}}(t) dt, \quad \theta_{\text{out}} = \int_0^T \Omega_{\text{out}}(t) dt \quad (9)$$

are kept the same. Under these conditions the only important parameter is the ratio $\theta_{\text{in}}/\theta_{\text{out}}$.

Fig. 2 Evolution of a blob originally placed at the position $x=2.2$ cm, $y=3.1$ cm (diameter 0.5 cm). Visualization is provided by a fluorescent tracer dissolved in glycerine excited by longwave ultraviolet light of 365 nm and is injected 2–5 mm below the free surface of the fluid by means of a syringe (the diffusion coefficient of the tracer in glycerine is $\sim 10^{-8}$ cm² s⁻¹). The top and bottom walls move according to the prescription of equation (5) with $U=2.69$ cm s⁻¹ ($Re=1.7$ and $0.05 < Sr < 0.09$). The pictures are taken at the instant when the top wall and bottom wall are moving at maximum and minimum (zero) speeds, respectively. The pictures correspond to different time periods T , a , $T=15$ s; b , $T=20$ s; c , $T=25$ s; d , $T=30$ s, and to wall displacements of 1,612 cm (a); 752 cm (b); 951 cm (c); 1,452 cm (d). In terms of qualitative appearance, there seems to be an optimum mixing at $T=20$ s. Beyond that, a large regular region seems to grow with increasing T . Note that the folded structure leaves holes, such as A; they are the result of incomplete horseshoes. The companion computer simulations are considerably more difficult, due to the fact that numerical diffusion is greater than the molecular diffusion in the fluid experiments.

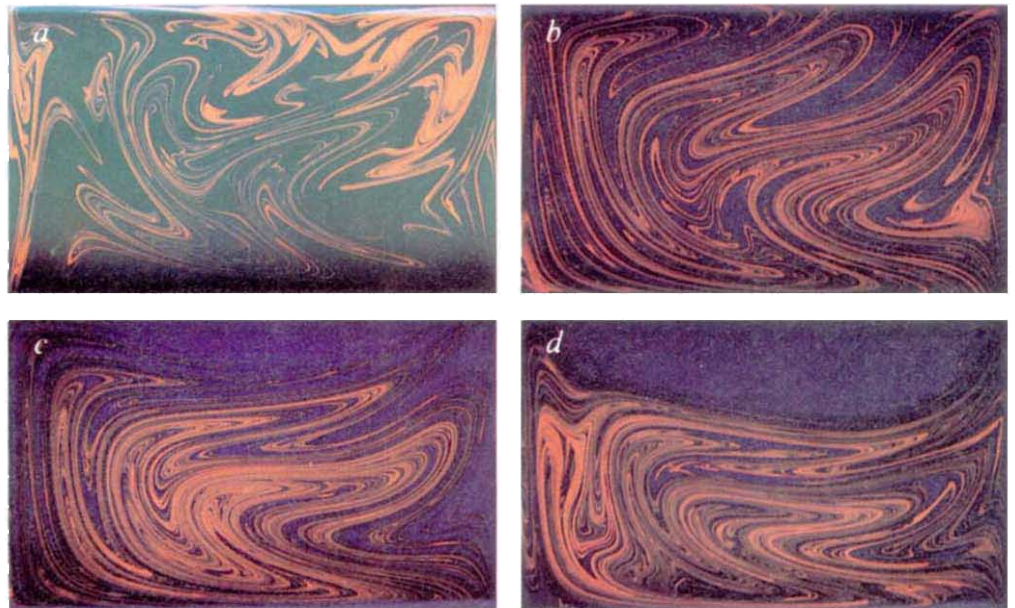


Fig. 3 Similar to Fig. 2 ($Re=1.2$, $0.08 < Sr < 0.13$) except that now the motion of the upper and lower walls is discontinuous; the top wall moves from left to right for half a period and then stops, then the bottom wall moves from right to left for half a period and then stops, and so on. In the experiments it takes some time for the motion to set in, $H^2/\nu \approx 5 \times 10^{-2}$ s. To minimize transient effects we wait for 5 s between the motion of the upper and lower bands. The pictures are taken at the end of the desired period. The periods are: a , $T=17$ s; b , $T=17$ s; c , $T=23$ s; d , $T=29$ s, and the corresponding displacements, 1,290 cm, 650 cm, 760 cm and 1,100 cm, respectively. In this case, the size of the regular (hole) region decreases as T increases and the best mixing is $T=29$ s. In a the blob of passive tracer was placed inside the island; in b the blob was placed outside the island; in c the lump structure surrounding the central island is clearly shown.

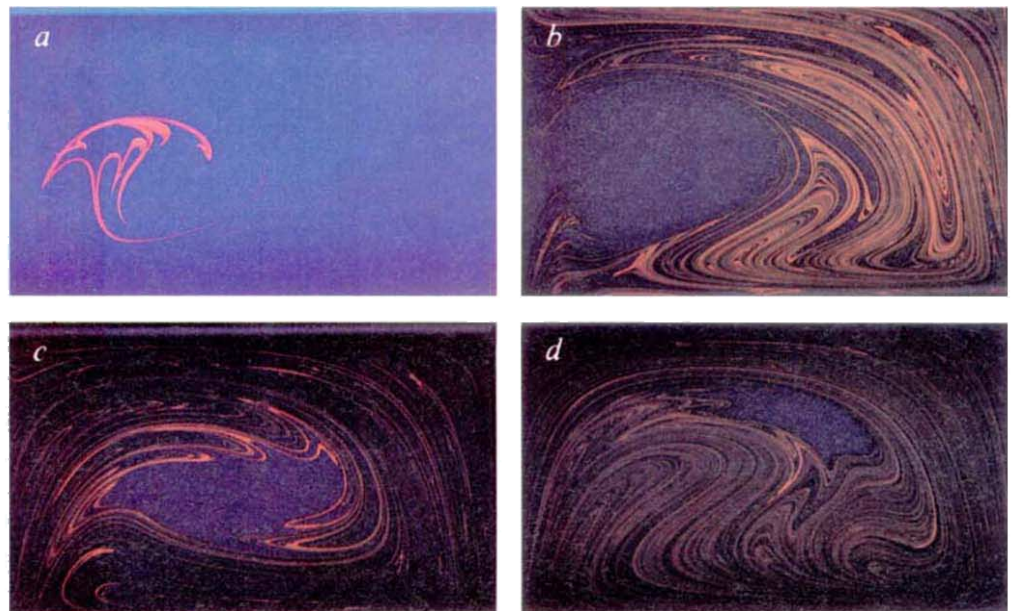
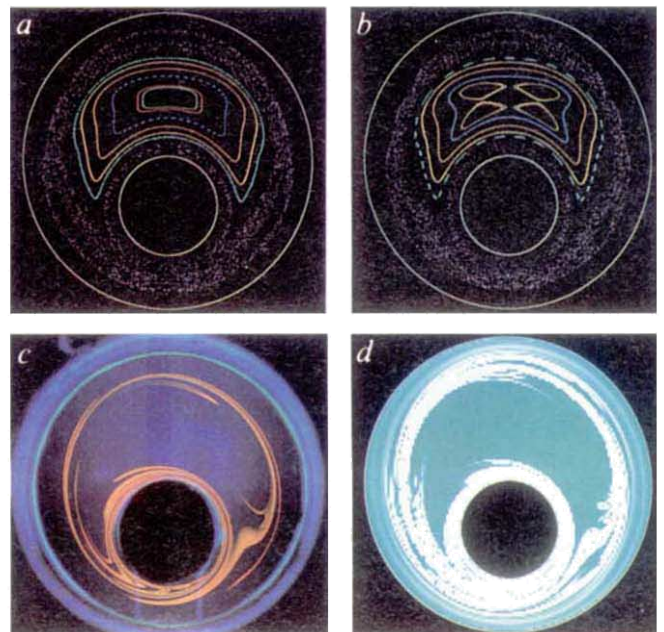


Fig. 4 Experimental and computational studies in the journal-bearing flow. The parameters values are $R_{in}/R_{out}=1/3$, $\delta/R_{out}=0.3$, and $\Omega_{in}/\Omega_{out}=-2$. a , Poincaré section with $\theta_{out}=165^\circ$; b , Poincaré section with $\theta_{out}=166^\circ$ (in both cases the computations correspond to 1,000 periods and 7 initial conditions); c , stretching of a blob initially located in the neighbourhood of the period-1 hyperbolic point; the stretching corresponds to 9 periods (the values of Re and Sr are 0.7 and 0.65, respectively, $Re=\rho V L/\nu$, and $Sr=L/TV$, where V is characteristic velocity $V=|V_{in}|+|V_{out}|$, $V_{in}=\Omega_{in}$, $V_{out}=\Omega_{out}$, L a characteristic length $L=R_{in}-R_{out}$, and T is the period of action of the boundaries) and d , stretching map corresponding to 10 periods and 200 initial orientations per pixel. The white regions correspond to stretching greater than 50.



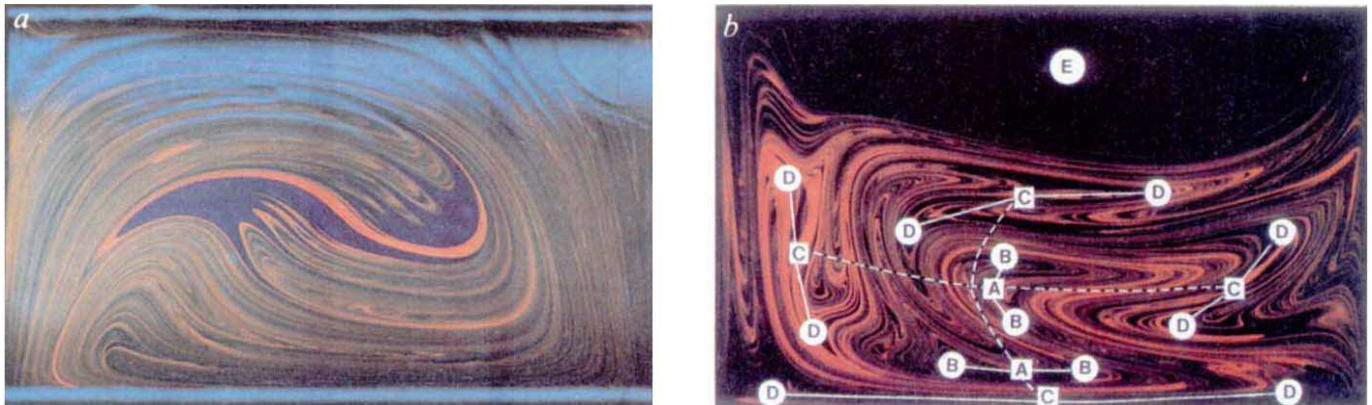


Fig. 5 *a*, An experiment similar to that of Fig. 3 showing the structure of an island at the point of bifurcation ($T = 24.3$ s, displacement = 970 cm, $Re = 1.2$, $Sr = 0.08$). This structure is characteristic of a golden-mean rotation speed. *b*, Partial structure of periodic points corresponding to the system of Fig. 2*d*. Key: A, hyperbolic period-1; B, elliptic period-2; C, hyperbolic period-2; D, elliptic period-4; E, hole of period-1. Note that the circles represent elliptic points and the squares hyperbolic points. The solid white lines connect two elliptic points to a central hyperbolic point; all three points move as a unit. The dotted lines connect two period-2 hyperbolic points to a period-1 hyperbolic point; the two period-2 hyperbolic points were born from the period-1 hyperbolic point. The two period-2 hyperbolic points interchange their positions after one period.

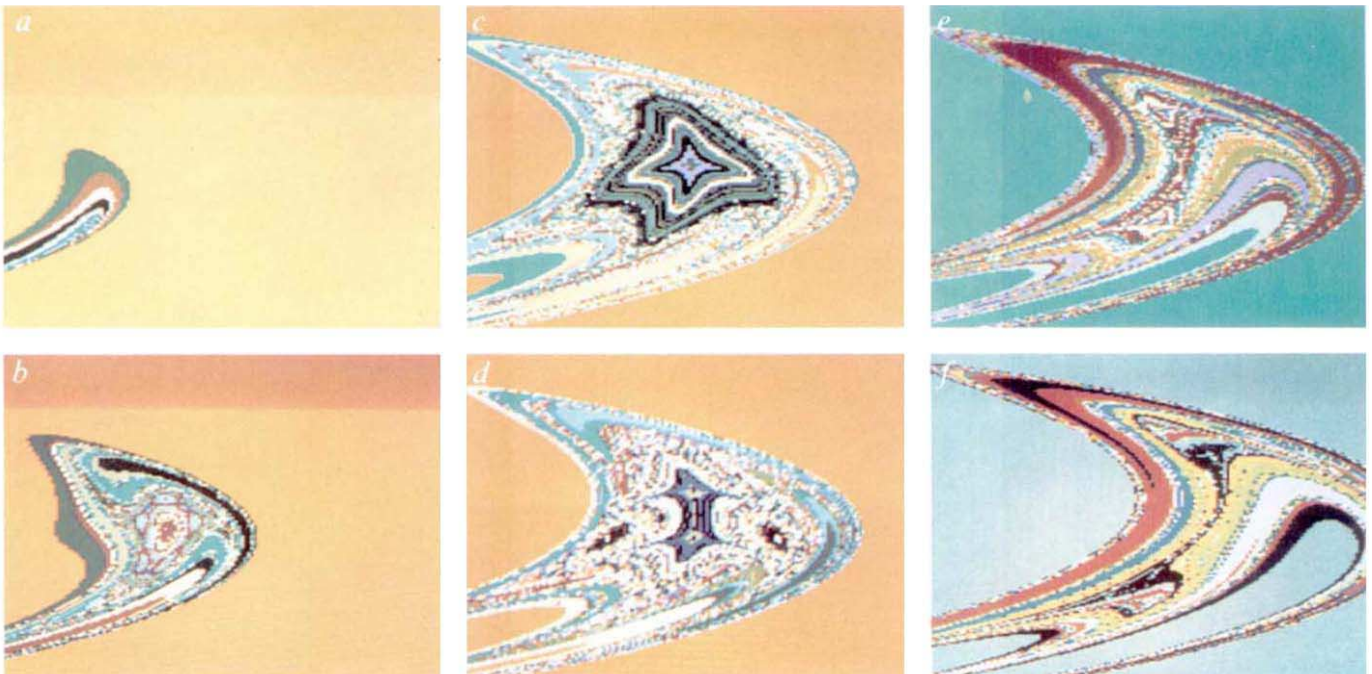


Fig. 6 Major structures observed during bifurcation sequence in mapping (11). The pictures displayed here are selected from a compressed screen animation of the bifurcation obtained by changing A (see mapping (11)), chosen to show the primary changes which occur. Points are coloured by number of iterations in an exponential manner and fall into three major colour groups. Unbounded points are coloured according to the number of iterations at which they exceed a default value. Bounded points are coloured by periodic approximation, determined by the point coming twice within a 'hit accuracy' of the reflection curves. The number of half-transformations between the two hits would be the period of the point if it actually landed on the curves (the mapping of equation (11) is the composition of two reflections; each reflection is known as a half-transformation). Points which do not fall into either group reach maximum iterations and are given the appropriate colour. The screens have maximum iteration value of 300 except *c*, for which this value is 800. The hit accuracy is 0.005 for 300 iterations and 0.001 for 800 iterations. The default value is -2 throughout. *a*, $A = -0.47$, shows the initial bifurcation to two fixed points, one hyperbolic, one elliptic, caused by the fluid motion folding over to cause a stagnation point. *b*, $A = -0.252$, shows that the elliptic point forms an island structure which grows in size, shedding rings of horseshoes and islands as the rotation speed of the boundary passes through rational rotation numbers. *c*, $A = 0$, at this point the elliptic point becomes hyperbolic in a 'period doubling' manner. Note that this does not imply the breakdown of the island (which is determined by the rotation speed at the edge, not the centre), and that the periodic point may revert to being elliptic without causing the destruction of the period-2 islands created in the event. *d*, $A = 0.117$, shows that the period-2 planetary islands formed in the period-doubling move into the surrounding flow. They in turn undergo a sequence of period doublings before being destroyed. *e*, $A = 0.238$, shows the appearance of a second set of period-2 points due to a second period-doubling bifurcation. For this particular system the island does break up at this point, the rotation speed of $1/2$ having penetrated the invariant rings of the centre of the island, *f*, $A = 0.241$. At this moment the remnants of the island can no longer be seen at this resolution.

Results

Figure 2 shows snapshots of the evolution of a blob in the cavity flow when the top and bottom walls move according to the $\sin^2$ flow of equation (5). The experiments suggest that the initial folding serves as a template for further stretching (image analysis suggests that the perimeter of the blob grows exponentially with the number of periods), and that the structure of islands shown in the photographs is very similar to the one observed for long times. The experiments indicate also that the holes (or islands, areas that the tracer has been unable to penetrate) and also the large-scale folds form 'coherent regions' that can be identified and followed in space and time; they are translated and deformed by the flow but conserve identity and do not disappear upon further mixing. It is obvious that the large-scale structures of all the figures is similar (see Fig. 2).

Experiments provide an indication of the nature and location of the boundary of the islands. The structure of the boundary is seen in Fig. 2c, d but a clearer indication is provided by Fig. 3. This shows similar results to Fig. 2 except that the motion of the upper and lower walls is discontinuous. The experiment, corresponding to a period of 17 s, provides a clear demonstration of both mixing within an island and the structure of the boundary. In Fig. 3a the blob of passive tracer was placed inside the island; in b the blob was placed outside the island. The boundary of the island is approached on either side, by 'folds' or 'lumps' corresponding to nearby manifolds (the points on the boundary are accumulation points of periodic points, transverse homoclinic and heteroclinic points, and homoclinic and heteroclinic tangencies¹⁷). This gives an estimate of the extent of the island. Parts of the interior of the lumps are in fact images of each other (they are portions of the same blob¹⁸). A careful counting of the lumps observed in the experiments (for example, see Fig. 3c), as well as observation of which lumps transform into each other in how many periods (using video recording), gives us an estimation of the rotational speed of the boundary (further details appear in ref. 19). It is clear that the island does not mix with rest of the fluid and that the stretching is small even though the displacement corresponding to Fig. 3a is roughly twice that of Fig. 3b. In this case, the size of the regular (hole) region decreases as T increases and the best mixing occurs at $T = 29$ s. It thus appears that the two prescriptions, the discontinuous protocol and the $\sin^2$ protocol, give opposite results.

However, this is only partially correct and this is one of the instances where computer modelling can help. Extensive computations carried out with the journal-bearing flow suggest that, as long as the displacements are kept the same, the structure of the system is remarkably similar for different protocols of motion of the boundary (even if $\Omega_{\text{in}}(t) \neq 0$ when $\Omega_{\text{out}}(t) \neq 0$). Suitably constructed Poincaré sections (preserving symmetry by starting the motion with half the displacement) appear to be insensitive to the flow prescription and depend mostly on the period rather than the shape of the motion of the wall; in particular the discontinuous and $\sin^2$ histories give nearly identical results. These conjectures are verified for the case of the cavity flow, and indeed if Figs 2 and 3 are compared at equal displacements of the walls a good agreement, in terms of the similarity of large-scale structures, is seen, especially at low values of T . Both systems, however, behave differently at large values of T .

This seems to be a strong point in favour of numerical studies but there is a danger of misinterpretation if analyses are based solely on a computational viewpoint. Figure 4a, b shows two Poincaré sections, with $\theta_{\text{in}}/\theta_{\text{out}} = -2$, but with $\theta_{\text{out}} = 165^\circ$ in a and with $\theta_{\text{out}} = 166^\circ$ in b (all other parameters are the same). A naive interpretation of these results, might suggest that the mixing structure in both systems should be very different as the structure of the islands is very different. Furthermore, if this were the case, the results would also suggest that extreme care should be exercised in experimental work because an error as little as 0.6% could affect the outcome of the mixing structure obtained. However, a deeper analysis reveals that this is not so

and these variations are unlikely to be of any consequence in the experiments. Indeed, extensive experiments, both in the cavity flow and the journal-bearing flow, show that often the rotation within large islands is nearly a solid-body rotation and 'whorls' are not noticeable even for a large number of iterations (20 or so). Analysis shows that the rate of rotation in the smaller islands is even slower and for all practical purposes they can be ignored in the stretching of lines places in this region.

However, the comparison between experiments and computations need not be of a direct nature; successful understanding can be achieved by making indirect connections. Experiments and computations for the system shown in Fig. 4c indicate that most of the stretching and folding is dominated by the unstable manifolds of hyperbolic points of low order. Experiments suggest also that unstable manifolds of low order capture blobs placed in the neighbourhood of the periodic point and provide a model for the large-scale structure of the system (Fig. 4c). It is therefore important to investigate whether or not the 'trapping regions' correspond also to the regions of highest stretch in the flow. To investigate the regions of significant stretching we adopt the following procedure: we integrate the equation

$$D\mathbf{F}/Dt = (\nabla \mathbf{v})^T \cdot \mathbf{F} \quad \text{with} \quad \mathbf{F}(\mathbf{X}, 0) = \mathbf{I} \quad (10)$$

where $\mathbf{F}$ is the deformation tensor ($\partial x_i / \partial X_j$), for the various initial conditions $\mathbf{X}$, and compute the stretching λ of a vector with initial orientation $\mathbf{M}$ as $\lambda = |\mathbf{F} \cdot \mathbf{M}|$. The average stretching is computed by averaging over many initial orientations. Figure 4d shows the results of such computations for a small number of periods. The white region corresponds to large stretching whereas the blue region corresponds to stretching of order one. The remarkable agreement of the computational and experimental results (Fig. 4c, d) is evidence that in this case the manifolds of the period-1 points provide a template for the stretching occurring in the flow and that most of the stretching is dominated by the period-one manifolds. Supplementary computations suggest the rate of spatial spreading is roughly proportional to the value of the eigenvalues and inversely proportional to the period of the point (that is, stretching experienced by a filament of fluid every time it passes in the neighbourhood of the periodic point).

The previous comparisons suggest that the relationship between computations and experiments is not trivial and indeed often misunderstood. In theory, there are always three systems; a mathematical system, a computer-simulated system, and an experimental system. However, it is naive to expect that all the information extracted from experiments is easily obtainable from computations, and vice versa. It is apparent that the requirements for computational observability are also different from those of experimental observability. In general, an object will be observable in computer experiments if it is robust, that is, relatively insensitive to round-off error, and also if it occurs with high probability in discrete floating-point calculations, that is, if its presence can be ascertained regardless of the pixel size on the computer screen. On the other hand, an experimental system is fundamentally analogue so that non-robust, low-probability events can be seen, although perhaps not easily repeated. Figure 5a shows an example of a bifurcation which would be hard to capture in feasible computer simulations. It shows the behaviour of the island boundary at the time of island collapse. This signature of lumps piled on each other is characteristic of the breaking of the most stable invariant boundary (boundary with rotation number = integer $\times (\sqrt{5} - 1)/2$, which is known as the golden mean). The coherence observed in these experiments can be exploited in various ways. Because it is possible to follow holes and folds in time and space it is possible to determine the location of the low-order periodic points. Figure 5b shows the partial structure of periodic points corresponding to the system of Fig. 2d along with a schematic diagram showing some of the bifurcations that take place as the period is increased. Such structure would be hard to simulate numerically.

Incomplete horseshoes

The experiments shown in Figs 2 and 3 reveal that by far the most common obstacles to complete mixing are incomplete horseshoes (indicated by A in the figures) rather than the conventional elliptical island of near-integrable systems of the KAM theorem. It is therefore important to investigate computer models capable of generating similar structures as well as capturing some of the essential aspects of the breakup of islands. (The term 'incomplete horseshoe' is used to denote the intermediate states in a bifurcation sequence from a saddle node to a complete Smale horseshoe.) (See Fig. 2.4.3 in ref. 9.) In the intermediate state, folds do not cross the quadrilateral defining the horseshoe. If the folds are pushed forward a complete horseshoe results. On the other hand, if the completed folds are pushed backwards there is a saddle-node collapse). The smoothness and regularity shown in the experiments are hard to achieve in computer simulations; finite pixel size (resolution) and round-off error govern the process and the regions often degenerate into pixel clouds by which the underlying structure is obscured (computer estimates are provided in ref. 20). Fortunately, the typical structures in Figs 2 and 3 are present in simple mappings amenable to detailed computer analysis. A model containing the essential behaviour of what might be termed 'blinking systems' (the blinking vortex^{8,9}, tendril-whorl mapping⁸, and in particular, the cavity flow and the journal-bearing flow) is the one arising from reflections of the plane through two parabolic conjugate lines²¹,

$$x_{n+1} = 2f(y_n) - x_n, \quad y_{n+1} = 2g(x_{n+1}) - y_n \quad (11)$$

(such a map is a representation of the flow in the neighbourhood of the intersection of two conjugate lines; the crossings determine the location of period-1 points; see ref. 3). The simplest system of this type is when the functions f and g are parabolas. Here we take $f(y) = A - y^2$ and $g(x) = -x^2$. Because the mapping consists of two reflections, it can be iterated at roughly the square of the speed of the other transformations thus allowing a thorough check of hypotheses about changes in the system with respect to changes in governing parameters. Figure 6 shows a computer simulation obtained using the map described by equation (11). Analysis reveals that the holes or missed folds labelled A in Fig. 2b correspond to saddle-node behaviour⁹ and that the severe folding of elliptical islands into s-shapes indicates a golden-mean breakup near either a period doubling, or the collapse of two saddle-node systems of the same period into two final periodic points. The folded structures at the boundary

of the island indicate the break-up of boundaries in the manner variously described by Takens²², Greene²³, Mather²⁴, and others (note that the boundary of the island is *not* a KAM curve and that in general is nowhere differentiable). However, the exact details of the process are an open and unexpectedly deep problem in hamiltonian dynamics and are not known at this time.

Applications

These type of studies can serve as useful prototypes for various problems in nature and technology spanning several orders of magnitude in length and timescales. For example, this kind of idealization can help in understanding the complex structures which appear during mixing in very low Reynolds number flows such as in structure formation in the Earth's mantle^{25,26} (even though these flows are not globally periodic), or the coherence and persistence of structures in two-dimensional flows such as mixing in oceans^{22,28} (even though the flows are not slow). The effectiveness of time periodicity in improving mixing can explain the role of periodic contractions of mixing in the duodenum²⁹, and suggest new engineering applications such as configurations for the processing of very viscous liquids such as polymers^{30,31} and shear sensitive biomaterials³². On a more speculative level, the problems discussed here might serve as a stepping stone for the understanding of turbulence (it has been conjectured that chaos might be a precursor to fully developed turbulence) and as a prototype for coherent structures in turbulent flows³³. However, from a fundamental viewpoint another important fact is that the equations for the fluid particle trajectories in the x - y plane provide a visualization of a hamiltonian system thereby providing an experimentally observable picture of phase space. In this context it is interesting to note that the visualization of hamiltonian systems by means of a fluid mixing (thought) experiment was noted a hundred years ago by J. Willard Gibbs³⁴. This transparent visual analogue readily distinguishes the phenomenon of chaotic trajectories in fluids from other studies in chaos, such as mappings in the complex plane³⁵ as the structures produced by fluid mixing are open to direct observation in terms of coherent or asymptotic structures detectable in conceptually simple cases.

We thank the NSF, the Department of Energy, the Air Force Office for Scientific Research and the Materials Research Laboratory of the University of Massachusetts for their support. H.R. would like to thank John Mather for clarification of the description of the boundary of islands. We would like also to thank J. G. Franjone for helpful comments on an earlier version of the manuscript.

Received 31 July 1987; accepted 13 March 1988.

1. Birkhoff, G. D. *Acta Math.* **43**, 1-119 (1920).
2. Moser, J. *Stable and Random Motions in Dynamical Systems* (Princeton University Press, 1973).
3. Khakhar, D. V., Rising, H. & Ottino, J. M. *J. Fluid Mech.* **172**, 419-451 (1986).
4. Chien, W.-L., Rising, H. & Ottino, J. M. *J. Fluid Mech.* **170**, 355-377 (1986).
5. Spencer, R. S. & Wiley, R. M. *J. Coll. Sci.* **6**, 133-145 (1951).
6. Arnold, V. I. *Mathematical Methods of Classical Mechanics* (Springer, New York, 1980).
7. Truesdell, C. & Toupin, R. A. in *Handbuch der Physik* vol. III/1 (ed. Flügge, S.) 226-793 (Springer, Berlin, 1960).
8. Aref, H. *J. Fluid Mech.* **143**, 1-21 (1984).
9. Guckenheimer, J. & Holmes, P. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields* (Springer, New York, 1983).
10. Lichtenberg, A. J. & Leibermann, M. A. *Regular and Stochastic Motion* (Springer, New York, 1982).
11. Hellemann, R. H. G. in *Fundamental Problems in Statistical Mechanics* (ed. Cohen, E. G. D.) 165-275 (North-Holland, Amsterdam, 1980).
12. Devaney, R. L. *An Introduction to Chaotic Dynamical Systems* (Benjamin/Cummings, Menlo Park, 1986).
13. Batchelor, G. K. *An Introduction to Fluid Mechanics* (Cambridge University Press, London, 1967).
14. Wannier, G. L. *Q. Appl. Math.* **VIII**, 1-32 (1950).
15. Chaiken, J., Chevray, R., Tabor, M. & Tan, Q. M. *Proc. R. Soc. A* **408**, 165-174 (1986).

16. Aref, H. & Balachandrar, S. *Phys. Fluids* **29**, 3515-3521 (1986).
17. Newhouse, S. E. *Institut des Hautes Etudes Scientifiques. Publications Mathématiques* **50**, 101-151 (1979).
18. Abraham, R. H. & Shaw, C. D. *Dynamics-The Geometry of Behavior* Pt 3, Fig. 6, 5.2.11 (Aerial Press, Santa Cruz, 1985).
19. Rising, H. thesis, Univ. Massachusetts (1988).
20. Franjone, J. G. & Ottino, J. M. *Phys. Fluids* **30**, 3641-3643 (1987).
21. MacKay, R. S. thesis, Princeton Univ. (1982).
22. Takens, F. *Math. Ann.* **188**, 304-312 (1970).
23. Greene, J. *J. math. Phys.* **20**, 1183-1201 (1979).
24. Mather, J. N. *Institut des Hautes Etudes Scientifiques. Publications Mathématiques* **63**, 153-204 (1986).
25. Allègre, C. J. & Turcotte, D. L. *Science* **323**, 123-127 (1986).
26. Hoffman, N. R. A. & MacKenzie, D. P. *Geophys. J. R. astr. Soc.* **82**, 163-206 (1985).
27. Veronis, G. *Adv. appl. Mech.* **13**, 1-92 (1973).
28. Rhines, P. B. A. *Rev. Fluid Mech.* **18**, 433-497 (1983).
29. Macagno, E. O. & Cristensen, J. A. *Reus. Fluid Mech.* **12**, 139-158 (1980).
30. Ottino, J. M. & Chella, R. *Polym. Eng. Sci.* **23**, 357-379 (1983).
31. Khakhar, D. V., Franjone, J. G. & Ottino, J. M. *Chem. Engng. Sci.* **42**, 2909-2926 (1987).
32. Sobey, I. C. *Chem. Engng Sci.* **40**, 2129-2134 (1985).
33. Roshko, A. *Am. Inst. Aeron. Astron.* **14**, 1349-1357 (1976).
34. Gibbs, J. W. *The Collected Works of J. W. Gibbs* Vol. II, Pt 1, 141-156 (Yale University Press, New Haven, 1948).
35. Devaney, R. L. *Science* **235**, 342-345 (1987).

Structure of the influenza virus haemagglutinin complexed with its receptor, sialic acid

W. Weis^{*,‡}, J. H. Brown^{*}, S. Cusack^{*,‡}, J. C. Paulson[†], J. J. Skehel[‡] & D. C. Wiley^{*,§}

^{*} Department of Biochemistry and Molecular Biology and Howard Hughes Medical Institute, Harvard University, Cambridge, Massachusetts 02138, USA

[†] Department of Biological Chemistry, UCLA School of Medicine, Los Angeles, California 90024, USA

[‡] National Institute for Medical Research, Mill Hill, London NW7 1AA, UK

The three-dimensional structures of influenza virus haemagglutinins complexed with cell receptor analogues show sialic acids bound to a pocket of conserved amino acids surrounded by antibody-binding sites. Sialic acid fills the conserved pocket, demonstrating that it is the influenza virus receptor. The proximity of the antibody-binding sites suggests that antibodies neutralize virus infectivity by preventing virus-to-cell binding. The structures suggest approaches to the design of anti-viral drugs that could block attachment of viruses to cells.

THE haemagglutinin glycoprotein (HA) of the influenza virus membrane is responsible for binding the virus to cell-surface receptors during infection. The only known components of the receptors necessary for this interaction are sialic acids (Fig. 1); the virus does not bind to or infect neuraminidase-treated cells^{1,2}, and binding is restored by re-sialation³ or addition of sialated glycolipid^{4,5}. The region of the haemagglutinin involved in receptor binding has been deduced from studies of mutant haemagglutinins with different binding specificities to involve a pocket of amino acids at its membrane distal surface⁶ (Fig. 2a). The pocket is surrounded by residues that vary with changes in antigenicity and is in fact the only surface exposed on the top of the HA that has not changed during the antigenic drift that has occurred in the HAs of Hong Kong influenza viruses since the pandemic of 1968 (Fig. 2b).

Using X-ray crystallography, we have determined the structure of the HA of a receptor binding mutant⁶ for comparison with the structure of the wild-type molecule³⁷, and we have also determined the structures of both mutant and wild-type HAs complexed with sialyllactose, a trisaccharide receptor analogue. The results obtained provide insight into how influenza virus attaches to a cell, they allow assessment of mechanisms of infectivity neutralization by antibodies, and they contain information which will assist development of anti-viral drugs designed to block viral entry into cells.

The receptor binding site

Topographically, the binding site is a depression, the bottom of which is formed by the phenolic hydroxyl of Tyr 98 and the aromatic ring of Trp 153 (Figs 3 and 2c). Glu 190 and Leu 194 project down from a short α -helix to define the rear of the site with His 183 and Thr 155 (Fig. 3). Residues 134 to 138 form the 'right' side of the site, and residues 224 to 228 form the 'left' side.

Interactions among the conserved residues forming the surface of the pocket, and between these and a conserved 'second shell' of residues, appear to orient several of the surface atoms for binding to ligand (Fig. 3): a chain of hydrogen bonds links Trp 153, Tyr 195, His 183, Tyr 98, and via a water molecule, Glu 190; Trp 153 is part of an edge-to-face stack of aromatic rings⁷ with Phe 147 and Phe (or Tyr) 148. This local stabilization may be important for maintaining the geometry of the site despite the numerous amino-acid substitutions observed to occur

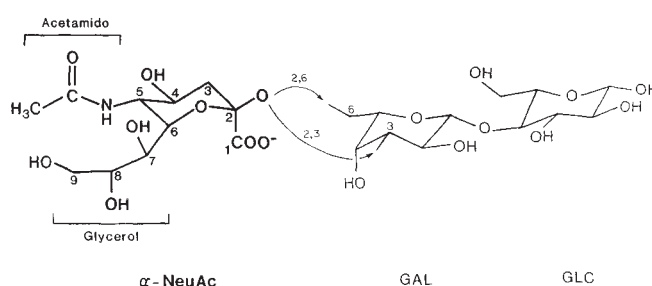


Fig. 1 Sialyllactose: α -N-acetyl neuraminic acid (α -NeuAc), a sialic acid, linked $\alpha(2,6)$ or $\alpha(2,3)$ to lactose (galactose- $\beta(1,4)$ -glucose), binds to the influenza HA. The α -anomers of sialic acids have their carboxylate axial to the pyranose ring at the 2 position. The acetamido and glycerol moieties of α -NeuAc are equatorial to the ring at the 5 and 6 positions, respectively.

around its perimeter in the haemagglutinins of new epidemic strains which arise every few years. Patterns of local stabilization in other saccharide-binding proteins have been discussed by Quijcho⁸.

HA-receptor analogue complexes

Complexes of wild-type HAs with $\alpha(2,6)$ linkage receptor specificity (Table 1 caption) were prepared by soaking HA crystals in $\alpha(2,6)$ sialyllactose (NeuAc $\alpha(2,6)$ Gal $\beta(1,4)$ Glc), a trisaccharide with the same composition and linkages as the termini of many cell surface oligosaccharides. Table 1 summarizes details of X-ray data collection and processing, and Fig. 4a shows the difference electron density between $\alpha(2,6)$ sialyllactose bound to HA, and HA without sialyllactose. The density shows a flat central core with in-plane protrusions corresponding to the ring substituents of α -NeuAc, and an out-of-plane bulge at one end protruding downward towards the protein that corresponds to the axial carboxylate substituent at C2. A similar difference electron density peak is seen between the complex of $\alpha(2,3)$ sialyllactose bound to the mutant HA and the mutant HA without sialyllactose (Fig. 4b), indicating that α -NeuAc is bound in essentially the same way in the two sites. Three water molecules found upon crystallographic refinement of both wild-type and mutant unliganded sites (Table 1) appear to be displaced by the α -NeuAc. The fact that the water molecules are near the carboxylate and glycerol 8 and 9 positions of α -NeuAc (Fig. 4c) appears to explain the weak or absent density around those atoms in Fig. 4a,b.

No electron density corresponding to the galactose or glucose moieties of sialyllactose can be seen in either complex. This suggests that in the crystals the lactose is spatially disordered

§ To whom correspondence should be addressed.

* Present addresses: EMBL, c/o I.L.L., 156x, 38042 Grenoble Cedex, France (S.C.); Department of Molecular Biophysics and Biochemistry and Howard Hughes Medical Institute, Yale University, 260 Whitney Avenue, New Haven, Connecticut 06511, USA (W.W.).

Table 1 Crystallographic data

Variant	Compound added	d_{lim} (Å)	a, c (Å)	Number of films (source)	R_{int} whole [¶]	R_{int} all(>%) [¶]	Unique reflections (% of total)	% > 3 σ	R factor [#]	r.m.s. deviations: 1–2 distances (Å) peptide torsions (°)	$R_{\Delta F}$ ^{**}
D1112G	$\alpha(2,6)$ sialyl- lactose*	3.2	162.6 177.4	32 (r.a.) [‡] 16 (CHESS)	0.135	0.160 (80)	63,089 (85)	64.7	0.204	0.023 3.5	0.168
G146D ⁵²	—	3.0	162.6 177.4	48 (r.a.)	0.115	0.115 (50)	74,765 (83)	71.4	0.209	0.023 3.3	0.169
L226Q	—	2.9	162.8 177.3	62 (r.a.) 6 (CHESS)	0.121	0.126 (70)	83,182 (83)	67.9	0.227	0.025 3.6	0.180
L226Q	$\alpha(2,3)$ sialyl- lactose [†]	2.9	162.8 176.8	32 (SSRL) 16 (CHESS)	0.114	0.125 (70)	76,438 (76)	70.8	0.222	0.024 3.4	

The bromelain-released haemagglutinins were prepared as previously described³⁷; all crystallize isomorphously in space group P4₁ (ref. 37). Crystals of D1112G were grown by microdialysis of a 35 mg ml⁻¹ protein solution against 1.30–1.34 M Na₃ citrate/0.1% NaN₃/pH 7.5. Crystals of L226Q were grown under similar conditions, except that citrate concentration was in the range 1.34–1.38 M. For the sialyllactose studies, crystals were washed by four serial transfers through 1.35 M Na₃ citrate/0.1% NaN₃/pH 7.5 (D1112G) or 1.48 M Na₃ citrate/0.1% NaN₃/pH 7.5 (L226Q) buffer to remove contaminating viral neuraminidase, then soaked in the same buffer containing 25 mM sialyllactose for 12 to 24 h before data collection. After data collection, the soaking solutions were assayed for free sialic acid³⁸ to ensure that residual neuraminidase had not broken down the substrate; no significant breakdown of sialyllactose occurred. Data were collected on 1° oscillation photographs, using both rotating anode (r.a.) and synchrotron X-ray sources. Films were scanned using the program SCAN12 (ref. 39), modified for the VAX by R. C. Ladner, and processed with programs provided by Dr. P. Evans of the MRC. Polarization corrections for the synchrotron films were done essentially as in Kahn *et al.*⁴⁰, using estimates of the fraction of the primary beam polarized in the horizontal direction (0.80 at SSRL, 0.82 at CHESS). Partially recorded reflections were corrected to their fully recorded equivalents using P. Evans' version of the postrefinement procedure⁴¹. Data sets were scaled to one another using a linear and an isotropic temperature (exponential) factor. All statistics given are for the range between 12.0 Å and the resolution limit of the data. Crystallographic least-squares refinement of the structures was performed with the program of Hendrickson and Konner⁴². Haemagglutinins are named relative to the A/Aichi/68 (H3N2) HA, using the single letter amino acid code. L226Q, for example, is a single amino acid mutant with 226 changed from L to Q: residues are numbered 1–328 in HA₁ and 1,001 to 1,175 in HA₂. D1112G and G146D both have wild-type receptor binding sites; their substitutions are remote from the site. They were studied in place of the wild type for practical reasons. The wild-type virus preferentially binds erythrocytes derivatized with $\alpha(2,6)$ linked sialic acid, while the L226Q mutant preferentially binds $\alpha(2,3)$ linked sialic acids².

* Isolated from human milk (Sigma; 85% $\alpha(2,6)$ isomer, 15% $\alpha(2,3)$).

† Isolated from bovine colostrum (Sigma; 85% $\alpha(2,3)$ isomer, 15% $\alpha(2,6)$).

‡ Rotating anode Elliot GX-6 at 39 kV × 19 mA or GX-13 at 39 kV × 39 mA, CuK α radiation, 100 μ m focus cup, Franks double mirror optics. Exposure time, 12 h per 1° film. Temperature, 4 °C.

§ Cornell High Energy Synchrotron Source. Wavelength, 1.557 Å (from postrefinement). Bent crystal monochromator (Si[111]) horizontal focus, mirror vertical. Exposure time, 40–80 s per 1°. Temperature, 20 °C.

|| Stanford Synchrotron Radiation Laboratory. Wavelength, 1.5418 Å. Optics as CHESS. Exposure time, 3–6 min per 1°. Temperature, 20 °C.

¶ $R_{\text{int}} = \sum_{hkl} \sum_{\text{obs}} |I_{\text{obs}} - \langle I_{\text{hkl}} \rangle| / \sum_{hkl} \sum_{\text{obs}} I_{\text{obs}}$. The $R_{\text{int,whole}}$ column gives the R factor on fully recorded intensities before postrefinement, and the $R_{\text{int,all}}$ column gives the R factor on intensities after postrefinement, including all reflections with percent recorded greater than the value given in parentheses.

$R = \sum_{hkl} |F_{\text{obs}} - F_{\text{calc}}| / \sum_{hkl} F_{\text{obs}}$.

** $R_{\Delta F} = \sum_{hkl} |F_{\text{obs}} - F_{\text{calc}}| / \sum_{hkl} F_{\text{obs}}$, where 1 and 2 refer to the data sets on the lines immediately above and below the entry in this column.

when bound, an observation consistent with binding studies done in solution (see below). After crystallographic least-squares refinement, only the C6' and C5' atoms linking α -NeuAc to Gal could be placed in the $\alpha(2,6)$ sialyllactose complex (that is, in 2F_o – F_c maps).

Neither difference Fourier map shows any evidence of conformational changes in the protein upon ligand binding.

Atomic model

We present an atomic model for how the HAs of a wild-type (Fig. 4c) and a mutant (Fig. 2c) influenza virus bind sialic acid, based on interatomic distances (Fig. 5) obtained from restrained least-squares refinement of the four structures listed in Table 1. The features of the model are found in both the wild-type and mutant sialyllactose complexes, but at the 3 Å resolution limit imposed by the HA crystals, it is not possible to define with complete certainty all hydrogen bonds or other intermolecular interactions. A discussion of supporting evidence for the model is given after a brief description.

The essential features of the model are that sialic acid is bound in the site with one face of the pyranose ring towards the bottom of the depression, and with the other face exposed to solution. Each of the ring substituents unique to α -NeuAc (Fig. 1) interacts with the protein, while the ring atoms and the 4-hydroxyl do not. One carboxylate oxygen, the acetamido

nitrogen, and the 8- and 9-hydroxyls of the glycerol face into the site and hydrogen bond with conserved side chain and main chain polar atoms (Fig. 5a, b); the acetamido methyl group is centred over and in van der Waals contact with the six-membered ring of conserved Trp 153; Trp 153 and conserved Gly 134, Leu 194 and His 183 form a non-polar surface complementary to and in contact with a non-polar surface on sialic acid formed by C9, C7 and the acetamido methyl group. The hydroxyls at positions 4 and 7 (Fig. 1) and the acetamido carbonyl oxygen face toward solution; the 7-hydroxyl and acetamido carbonyl may form an intramolecular hydrogen bond as observed previously in crystals of the β -NeuAc methyl ester⁹.

About 66% (310 Å²) of the solvent-accessible surface¹⁰ of the sialic acid is buried upon binding to the HA, and all of the atoms making hydrogen bonds to the protein (the 8 and 9 hydroxyls, one carboxylate oxygen, and the acetamido nitrogen) are completely inaccessible to solvent when bound. This suggests that these atoms form stronger hydrogen bonds to the protein than to water, as regions isolated from bulk solvent tend to have smaller dielectric constants¹¹. The interaction of the carboxylate with a main chain amide proton (N137) rather than with a positively charged residue appears to be another example of the 'solvation' of buried charges by peptide bonds¹², although the partial solvent accessibility of the other carboxylate oxygen may diminish this effect by dispersing the charge.

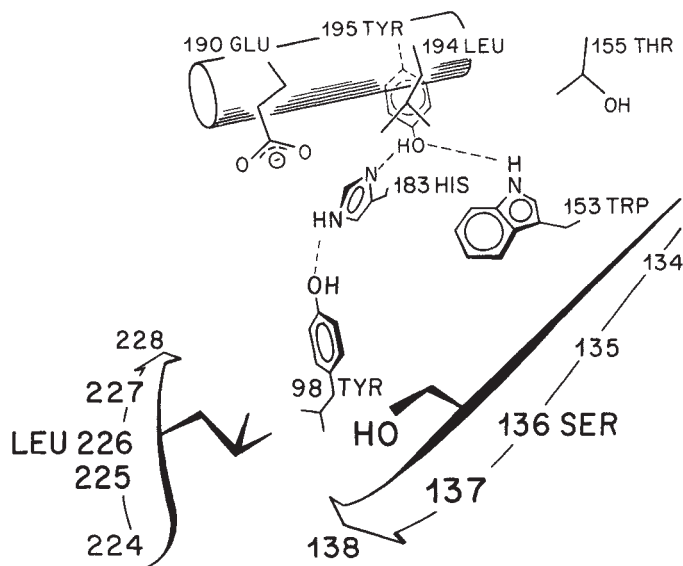


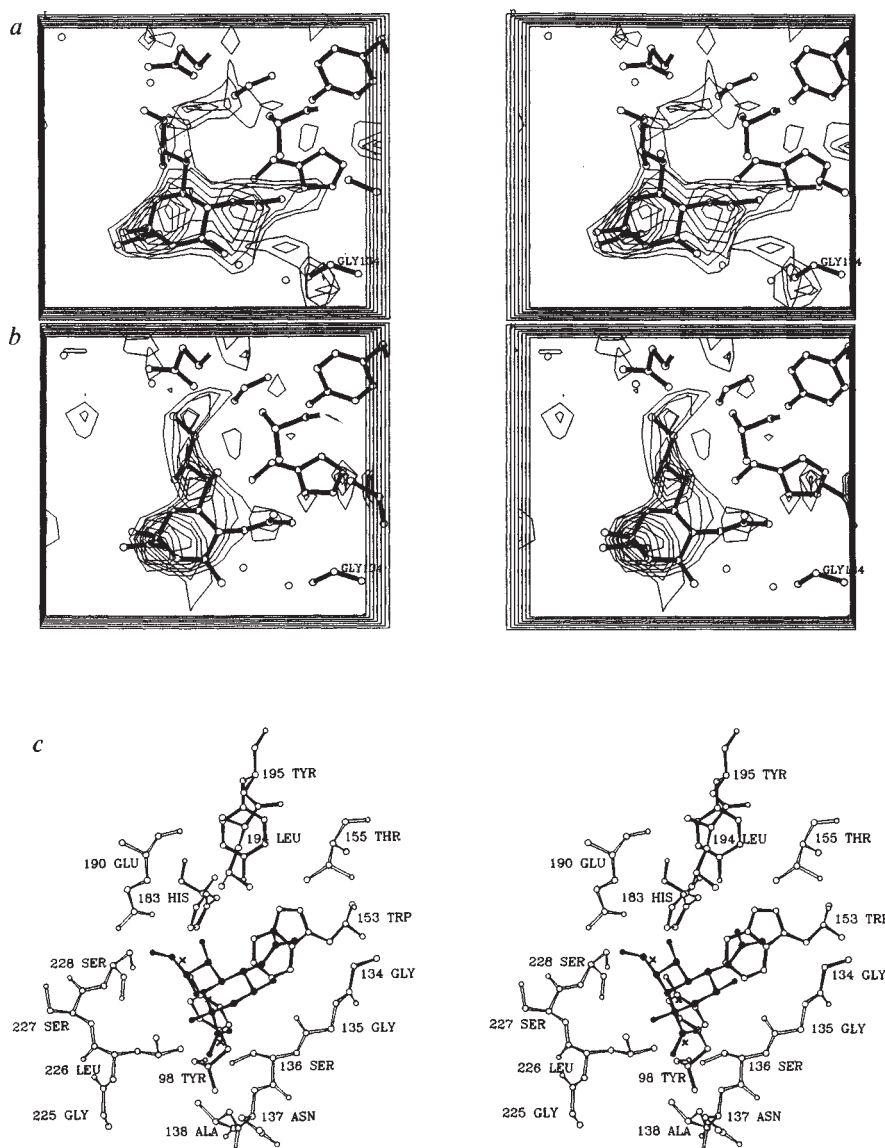
Fig. 3 The receptor binding site showing hydrogen bonding among conserved residues in site. See text for details.

Some of the observed interactions are similar to those in the recognition of *N*-acetyl groups by wheat germ agglutinin (WGA)^{13,14} and lysozyme¹⁵: in WGA the methyl group contacts the center of a tyrosine ring, and in lysozyme, an amide proton hydrogen bonds to a main chain carbonyl oxygen.

Mutant binding site

The amino acid substitution of glutamine for leucine at 226 (L226Q) reduces the affinity of the HA for sialosides with $\alpha(2,6)$ linkages relative to $\alpha(2,3)$ linkages^{6,16,17}. Recent proton nuclear magnetic resonance (NMR) measurements of the binding of $\alpha(2,6)$ and $\alpha(2,3)$ sialyllactose to the mutant HA (N. Sauter, M. Bednarski, B. Wurzburg, G. Whitesides, J. Skehel and D. Wiley, unpublished) indicate an affinity reduction of about two-fold (or only about 0.4 Kcal mol⁻¹ difference in binding energy). No direct X-ray data are currently available to help explain this reduction in affinity. In the wild-type plus $\alpha(2,6)$ sialyllactose and mutant plus $\alpha(2,3)$ sialyllactose complexes (which have approximately equal affinities in solution), however, none of the linkage atoms (O2 of NeuAc, C6', C3' of Gal) are near the side chain at position 226. Thus the affinity reduction does not appear to be determined by a direct interaction between the substituted side chain and the linkage atoms. Instead, the affinity difference is more likely to result from

Fig. 4 Stereo pairs of α -NeuAc fitted to difference electron-density maps of wild-type (a) and L226Q mutant (b) site. See text for interpretation. The α -NeuAc model shown is based on the crystal structure of the β -anomer⁴⁷ with the carboxylate moved to the axial position. Maps have been averaged about the molecular three-fold symmetry axis, using the programs of Bricogne⁴⁸. a, Difference density between wild-type site (D1112G) plus $\alpha(2,6)$ sialyllactose and the uncomplexed site. The peak shown is the only significant peak on the map aside from a peak at 1112 due to known sequence differences⁴⁹. (The map is calculated with coefficients ($F_{\text{obs}}^{\text{D1112G}+(2,6)\text{sialyllactose}} - F_{\text{obs}}^{\text{G146D}}$) and phases from a refined G146D model⁵⁰; contoured at 1.75 to 3.25 σ above the mean density, in steps of 0.5 σ .) b, Difference electron density between mutant L226Q binding site plus $\alpha(2,3)$ sialyllactose and the uncomplexed mutant site. The peak shown is one of two significant peaks on the map. The other is a spherical peak, smaller than a monosaccharide, located in an interior cavity on the molecular three-fold symmetry axis⁴⁹. (The map is calculated with coefficients ($F_{\text{obs}}^{\text{L226Q}+(2,3)\text{sialyllactose}} - F_{\text{obs}}^{\text{L226Q}}$). The phases are calculated from a model of L226Q based on the G146D model⁵⁰, refined for 13 cycles of restrained least squares refinement⁴² against the L226Q+sialyllactose data (Table 1). The map is contoured at 2.0 to 3.5 σ above the mean density, in steps of 0.5 σ .) c, Wild-type receptor site with sialic acid, stereo pair. Residues with solid bonds are conserved in all known HA sequences from natural isolates. Sialic acid is shown with solid spheres. Three water molecules displaced by sialic acid binding are indicated by X's. Criteria for water molecule positioning: (1) present in three-fold averaged and unaveraged $F_0 - F_c$ maps; (2) present in both wild type (G146D) and mutant (L226Q) $F_0 - F_c$ maps; (3) positions and temperature factors refined stably in both crystals; (4) contour level above any significant peaks in bulk solvent region; (5) reasonable distances to hydrogen bond donor and acceptors. Diagram drawn with ARPLOT⁵¹.



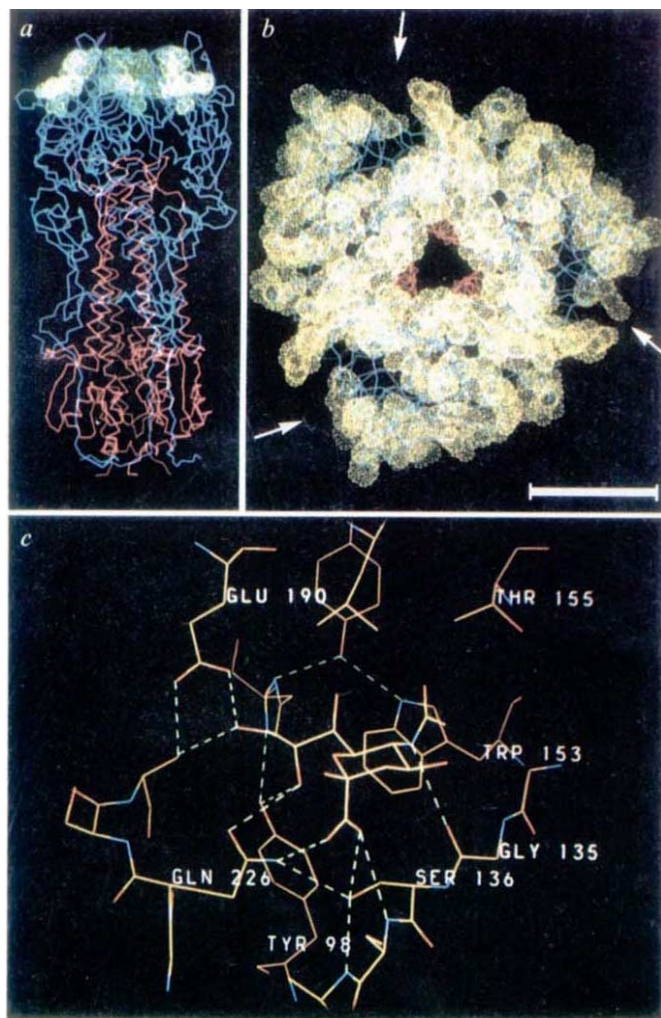


Fig. 2 The influenza virus haemagglutinin binding sites. *a*, Location of the three receptor binding sites on the influenza haemagglutinin. The HA is a trimer composed of three HA₁ chains (blue) and three HA₂ chains (red) (only α carbons are shown). Each site is composed of conserved residues within a single HA₁ monomer; these residues are shown as green dot surfaces. *b*, Top view of the haemagglutinin trimer illustrating antigenic variation around the three receptor-binding sites (white arrows). Only the α carbons (coloured as in *b*) are shown. Residues in yellow are those of the A/Aichi/68 HA that have varied between 1968 and 1982 in natural isolates, and also those from antigenically distinct viruses with single amino acid substitutions selected by growth in neutralizing monoclonal antibodies. The *N*-linked carbohydrate at Asn 165, which protects the protein surface from antibody pressure⁴³, is also shown in yellow. The bar represents 25 Å, the approximate size of an antibody 'footprint'^{25,26,27}. Antigenic sites surrounding receptor binding sites have been noted for influenza neuraminidase⁴⁴, influenza HA⁴⁵ and rhinovirus⁴⁶. *c*, Binding site of L226Q HA with α -NeuAc bound. Yellow, carbon; blue, nitrogen; red oxygen; dashed green lines, one possible set of hydrogen bonds (see text and Fig. 5*b*). Picture generated on Evans and Sutherland PS300 with HYDRA (R. Hubbard, unpublished).

conformational differences between the mutant and wild-type proteins.

Figure 6 shows the difference in conformation between the mutant L226Q and wild-type HAs in the binding site region, based on the analysis of a difference electron density map and refinement of both structures (Table 1). Positive and negative difference electron density peaks indicate that the carbonyl oxygen and amino nitrogen of the glutamine side chain project up toward solution above one leucine methyl position, whereas the position occupied by the 'lower' methyl of the leucine is vacated (Fig. 6). A series of other positive and negative difference electron density peaks indicate small adjustments (<1 Å) of residues in the site, presumably to fill the void created by the loss of one methyl group of Leu 226 and to accommodate the new potential for hydrogen bonding of the glutamine. Tyr 98 occupies some of the space vacated by this methyl group, and Trp 153, Phe 147 and His 183 adjust to accommodate this change. The loss of a contact between Leu 226 and Ala 138 (Fig. 5*a, b*) "allows" the site to 'close' slightly and a new hydrogen bond to be made across the site from Gln 226 to Ser 136. Smaller but significant shifts in residue position (not shown) appear to be transmitted through positions 185 (near His 183) and 228 to 218, 219, 222 on a β strand adjacent to 226, and to residues 246 and 165 (including its *N*-linked oligosaccharide) adjacent to this strand but across the trimer subunit boundary. The conformational adjustments seen in the mutant HA are consistent with earlier suggestions^{18–20} that changing residues remote from the site (244 adjacent to 246, 185, 188, 189, 218, 231 adjacent to 184) can affect receptor affinity. In some way these small conformational differences produce a reduced affinity for α (2,6) linked sialosides relative to α (2,3) sialosides.

Surprisingly, deletion of residues 224–230, which removes the left side of the site (Fig. 3) results in a viable virus, albeit with altered binding characteristics²¹. This deletion does not remove any of the conserved residues or atoms that interact directly with α (2,6) sialosides.

Receptor binding in solution

The interactions of each substituent of the sialic acid ring with the HA observed in the crystal structures are consistent with those deduced from measurements in solution. The acetamido methyl proton NMR resonance exhibits a strong upfield chemical shift upon binding to the HA, characteristic of an interaction with the face of an aromatic ring (N. Sauter, M. Bednarski, B. Wurzburg, G. Whitesides, J. Skehel and D. Wiley, unpublished). This is consistent with the observation (Fig. 5*a, b*) that the methyl group interacts with the center of the six-membered ring of Trp 153. This ring-current shift occurs in the binding of both α (2,6) and α (2,3) sialyllactose, as well as other sialosides, to HAs of both the wild-type and L226Q mutant viruses, confirming our observation (Fig. 4*a, b*) that sialic acid binds in the same orientation to both HAs. A change at amino acid 155 from Thr to Tyr has been correlated with an increase in affinity for *N*-glycolyl neuraminic acid, a sialic acid with the methyl group of the acetamido substituent replaced by a hydroxymethyl group^{22,23}. This is also consistent with the location of the methyl group over Trp 153 and facing towards amino acid 155 (Figs 3, 4*c*).

The interactions observed in the crystal structures at the glycosidic linkage, the carboxylate, and the 9-hydroxyl are consistent with relative binding affinities for the HA assessed by

the ability of several sialosides to delay viral adsorption to sparsely sialated erythrocytes¹⁷. Inhibition by α -anomeric sialosides is relatively independent of the group attached at the glycosidic oxygen (O2) of sialic acid (ref. 17 and T. Pritchett, R. Brossmer and J. Paulson, unpublished). This is consistent with the spatial disorder of lactose inferred from the crystal structures (Fig. 4a, b), which suggests that components of α -sialosides linked at O2 do not interact strongly with the HA and therefore contribute little to inhibition of viral adsorption. Evidence that the axial carboxylate at C2 of α -sialosides interacts with the protein in solution as observed in the crystal structures (Fig. 4c) is provided by the observations that α -anomers (axial carboxylate) inhibit 30-fold better than β -anomers (equatorial carboxylate), and that the methyl ester of α -anomers is a 10-fold weaker inhibitor than the free acid (T. Pritchett, R. Brossmer and J. Paulson, unpublished), consistent with the loss of a charged hydrogen bond and the fairly tight fit of the carboxylate in the site. The absence of any interaction between the 4-OH and the HA is consistent with the observation that acetylation at that position does not affect inhibition²⁴. Conversely, the tight packing of O9 seen in the crystal structures may explain why influenza A viruses fail to bind to erythrocytes derivatized exclusively with 9-O-acetyl neuraminic acids²²; the addition of an acetyl group of O9 would not appear to fit in the complexes that we have observed.

Neutralization of virus infectivity

In this study, the receptor binding site of the influenza virus HA has been shown to be a conserved pocket of amino acids surrounded by antigenically variable antibody binding sites. Sialic acid occupies the entire pocket, indicating that this monosaccharide is the dominant component of the influenza virus cellular receptor, and the only component for which a binding site is conserved during the antigenic evolution of the HA. These observations provide a structural proof that the influenza virus receptor is sialic acid.

The majority of antigenic sites surround the receptor-binding site, and it appears likely that antibodies bound to these sites would block receptor binding, just as *in vitro* they prevent haemagglutination. Furthermore, because of the large binding surface of an antibody (20 × 30 Å 'footprint'²⁵⁻²⁷), it is probable that even if an antibody were directed at the conserved residues in the sialic acid binding site, it would contact a number of non-conserved residues on the edges of the site and would therefore be rendered ineffective by the antibody-selected variation of those residues. Neither the precise mechanisms involved in the neutralization of virus infectivity by antibodies nor the number of antibodies required for neutralization either *in vitro* or *in vivo* have been directly determined (reviewed in refs 28-30). It is possible, for example, that virus-antibody complexes are not taken into endosomes even if limited HA-receptor interaction is sufficient for virus-cell association, and it is also possible that antibody binding to certain regions of the HA prevents the membrane fusion activity of the HA more effectively than it does receptor binding. From the results of this study, however, indicating the proximity of receptor binding and antibody binding sites, it seems reasonable to propose that a major mechanism by which antibodies neutralize influenza virus is by an effect on virus-receptor interactions.

Receptor tropism

The HA specificity for $\alpha(2,6)$ and $\alpha(2,3)$ linked sialosides of human, avian and equine influenza isolates correlates with species of origin, suggesting that receptor specificity plays a role in the biology of the virus¹⁶. Avian influenza viruses replicate in the intestinal tract of their host, and this tissue tropism appears to require Gln at position 226 and Gly at position 228 of the haemagglutinin³¹. The haemagglutinin of influenza virus A/Duck/Ukraine/63, a virus which may be closely related to the progenitor of the human Hong Kong influenza viruses³²,

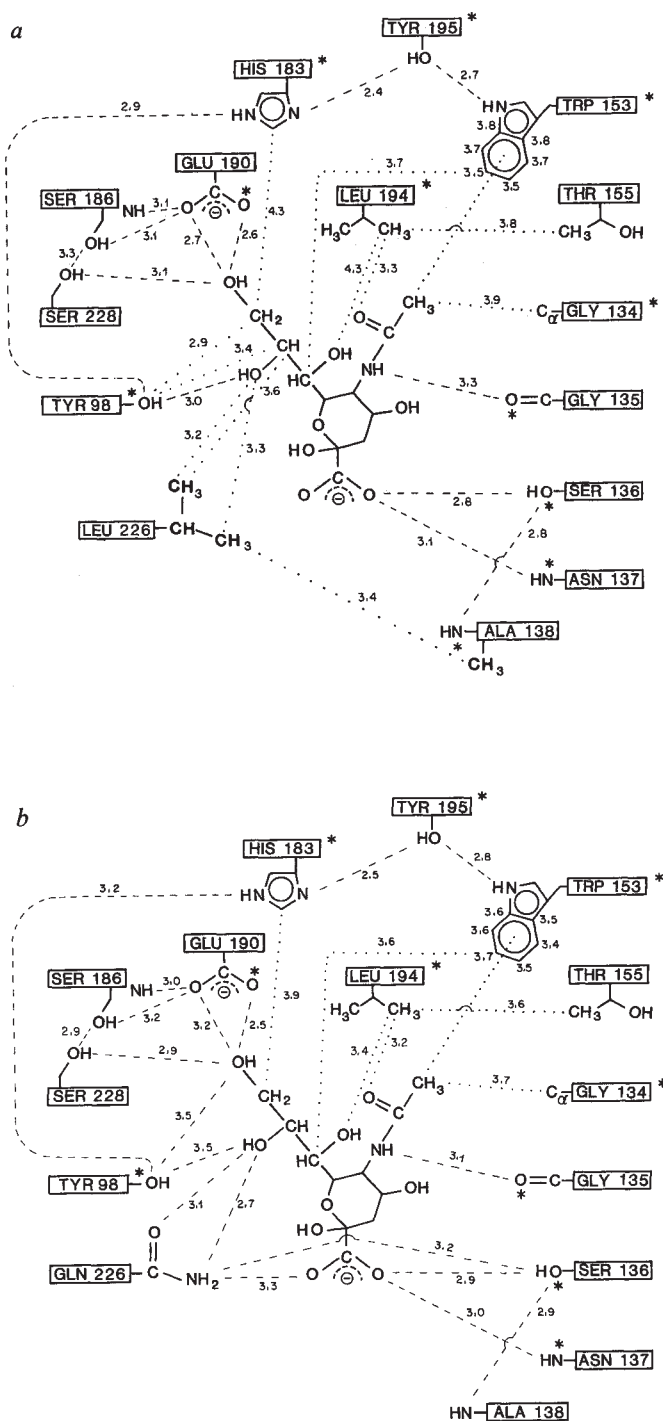
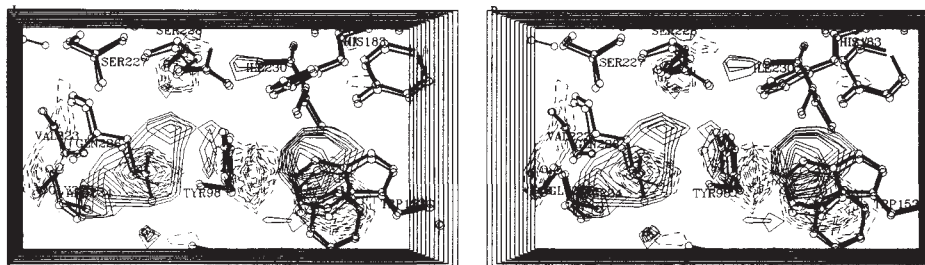


Fig. 5 Potential interactions of α -NeuAc with wild-type (a) and L226Q (b) virus HA. Potential hydrogen bonds (dashed lines) and van der Waals contacts (dotted lines) in wild type (D1112G) + $\alpha(2,6)$ sialyllactose (a) and in L226Q + $\alpha(2,3)$ sialyllactose (b). A star (*) next to a box indicates that the residue is conserved in all known HA sequences from natural isolates; a star next to an atom means that it is main chain or conserved in substitutions. a, Coordinates from a refined model of D1112G + α -NeuAc. b, Coordinates from a model of L226Q + α -NeuAc. (The distances shown (in Å) are averages of the same distance in the three sites of the refined trimer models. Distances shown next to the atoms of the six-membered ring of Trp 153 are to the acetamido methyl group of NeuAc. Some atoms are shown with more hydrogen bonds than they can actually participate in, due to uncertainties in the refined coordinates [the coordinate error is approximately 0.3 Å].)

Fig. 6 Stereo pair of the difference electron density between the mutant L226Q and wild-type binding sites. The refined G146D (wild-type) and L226Q (mutant) coordinates are superimposed, with the mutant in the bolder lines. (See text for interpretation.) (The map is calculated with coefficients $(F_{\text{obs}}^{\text{L226Q}} - F_{\text{obs}}^{\text{G146D}})$ and the same phases used in the map shown in Fig. 4a. Map contoured at 3.0–4.5 σ above the mean, in steps of 0.5 σ . Positive contours are shown as solid lines, negative dashed. Averaged about the molecular three-fold symmetry axis.)



differs from the HA of the 1968 Hong Kong virus by 21 amino acids^{33,34}. One of these differences is glutamine at position 226, and accordingly this avian virus, like other avian and equine influenza viruses^{16,35}, preferentially binds to $\alpha(2,3)$ linked sialoglycoconjugates³⁶. A mutant of A/Duck/Ukraine/63 selected for $\alpha(2,6)$ binding has the single substitution Q226L³⁶, the reverse of the receptor site mutant studied here, indicating that even in a background of other amino acid substitutions, the conformational differences that we have observed between the single amino acid mutant and wild-type HAs influence receptor binding specificity.

Prospects for drug design

We have presented an atomic model for how influenza virus recognizes its cellular receptor, sialic acid. The details of the sialic acid–HA interaction provide a possible basis for the design of anti-viral drugs that would block viral attachment to cells. An inhibitor targeted to the conserved residues and atoms of the receptor binding site is particularly attractive, as it might be effective against influenza virus of all subtypes. This is a comparatively simple system for drug design: upon binding, no chemical reaction takes place, and apparently no conformational changes occur in the protein. Moreover, a drug targeted at the receptor binding site needs only to be presented extracellularly (to the respiratory tract), avoiding the complication of creating a compound that can cross a cell membrane. At present,

however, our ability to design such drugs is limited by the resolution of our X-ray data, as well as the more fundamental problem of predicting the energetics of an intermolecular interaction from structural data.

Despite the limitations of the present structures, the interactions identified in this work lend themselves to design of HA inhibitors. Strengthening existing intermolecular hydrogen bonds by changing the electronic structure around the relevant ligand atoms may be possible, as well as adding additional atoms at favourable sites. Any attempt at drug design will have to consider the cooperativity of the virus–cell surface interaction. Even a compound with a substantially higher intrinsic affinity for the site will probably have to be presented as a polymer or on a microsurface to overcome the large effective affinity produced by the cooperative binding.

We thank Peter Rosenthal and David Stevens for excellent technical assistance, and Paul Phizackerley of SSRL and Keith Moffat and Wilfried Schildkamp of CHESS for assistance at those facilities. We acknowledge the contributions of Ian Wilson, Andy Cherenson, and Frank Escobar in our earlier studies of α -NeuAc–HA interactions. W.W. was supported by NSF and W. R. Grace predoctoral fellowships. S.C. was supported by an EMBL Visiting Fellowship, J.C.P. by the NIH and D.C.W. by grants from the NIH, NSF and synchrotron project CHESS and SSRL proposals.

- Gottschalk, A. in *The Viruses* Vol. 3 (eds Burnet, F. M. & Stanley, W. M.) 51–61 (Academic, New York, 1959).
- Paulson, J. C. in *The Receptors* Vol. 2 (ed. Conn, P. M.) 131–219 (Academic, Orlando, 1985).
- Paulson, J. C., Sadler, J. E. & Hill, R. L. *J. biol. Chem.* **254**, 2120–2124 (1979).
- Bergelson, L. D. *et al. Eur. J. Biochem.* **128**, 467–474 (1982).
- Suzuki, Y., Matsunaga, M. & Masumoto, M. *J. biol. Chem.* **260**, 1362–1365 (1985).
- Rogers, G. N. *et al. Nature* **304**, 76–79 (1983).
- Burley, S. K. & Petsko, G. A. *Science* **229**, 23–28 (1985).
- Quiocho, F. A. *Rev. Biochem.* **55**, 287–315 (1986).
- O'Connell, A. M. *Acta crystallogr.* **B29**, 2320–2328 (1973).
- Richards, F. M. *Rev. Biophys. Bioeng.* **6**, 151–176 (1977).
- Quiocho, F. A. in *Current Topics in Microbiology and Immunology* (in the press).
- Quiocho, F. A., Sack, J. S. & Vyas, N. K. *Nature* **329**, 561–564 (1987).
- Wright, C. S. *J. molec. Biol.* **178**, 91–104 (1984).
- Kronis, K. A. & Carver, J. P. *Biochemistry* **24**, 826–833 (1985).
- Kelly, J. A., Sielecki, A. R., Sykes, B. D., James, M. N. G. & Phillips, D. C. *Nature* **282**, 875–878 (1979).
- Rogers, G. N. & Paulson, J. C. *Virology* **127**, 361–373 (1983).
- Pritchett, T. J., Brossmer, R., Rose, U. & Paulson, J. C. *Virology*, **160**, 502–506 (1987).
- Yewdell, J. W., Caton, A. J. & Gerhard, W. *J. Virol.* **57**, 623–628 (1986).
- Underwood, P. A. *Arch. Virol.* **84**, 53–61 (1985).
- Underwood, P. A., Skehel, J. J. & Wiley, D. C. *J. Virol.* **61**, 206–208 (1987).
- Daniels, R. S. *et al. EMBO J.* **6**, 1459–1465 (1987).
- Higa, H. H., Rogers, G. N. & Paulson, J. C. *Virology* **144**, 279–282 (1985).
- Anders, E. M., Scalzo, A. A., Rogers, G. N. & White, D. O. *J. Virol.* **60**, 476–482 (1986).
- Pritchett, T. J. thesis, Univ. of Calif., Los Angeles (1987).
- Amit, A. G., Mariuzza, R. A., Phillips, S. E. V. & Poljak, R. J. *Science* **233**, 747–753 (1986).
- Colman, P. M. *et al. Nature* **326**, 358–363 (1987).
- Sheriff, S. *et al. Proc. natn. Acad. Sci. U.S.A.* **84**, 8075–8079 (1987).
- Mandel, B. *Comprehensive Virology* **15**, 37–121 (eds Fraenkel-Conrat, H. & Wagner, R. R.) (Plenum, New York, 1979).
- Dimmock, N. J. *J. gen. Virol.* **65**, 1015–1022 (1984).
- Dimmock, N. J. *Trends biochem. Sci.* **12**, 70–75 (1987).
- Naeye, C. W., Hinshaw, V. S. & Webster, R. G. *J. Virol.* **51**, 567–569 (1984).
- Laver, W. G. & Webster, R. G. *Virology* **51**, 383–391 (1973).
- Fang, R., Minjou, W., Huylebroeck, D., Devos, R. & Fiers, W. *Cell* **25**, 315–323 (1981).
- Ward, C. W. & Doppeide, T. A. *Biochem. J.* **195**, 337–340 (1981).
- Daniels, R. S., Skehel, J. J. & Wiley, D. C. *J. gen. Virol.* **66**, 457–464 (1985).
- Rogers, G. N. *et al. J. biol. Chem.* **260**, 7362–7367 (1985).
- Wilson, I. A., Skehel, J. J. & Wiley, D. C. *Nature* **289**, 366–373 (1981).
- Warren, L. *J. biol. Chem.* **234**, 1971–1975 (1959).
- Crawford, J. L. Ph. D. thesis, Harvard University (1977).
- Kahn, R. *et al. J. appl. Cryst.* **15**, 330–337 (1982).
- Wiukler, F. K., Schutt, C. E. & Harrison, S. C. *Acta crystallogr.* **A35**, 901–911 (1979).
- Hendrikson, W. A. & Konnert, J. H. in *Biomolecular Structure, Function, Conformation and Evolution* (ed. Srinivasan, R.), Vol. 1, 43–57 (Pergamon, Oxford, 1981).
- Skehel, J. J. *et al. Proc. natn. Acad. Sci. U.S.A.* **81**, 1779–1783 (1984).
- Colman, P. M., Varghese, J. N. & Laver, W. G. *Nature* **303**, 41–44 (1983).
- Wiley, D. C., Wilson, I. A. & Skehel, J. J. *Nature* **289**, 373–378 (1981).
- Rossmann, M. G. *et al. Nature* **317**, 145–153 (1985).
- Flippin, J. L. *Acta crystallogr.* **B29**, 1881–1886 (1973).
- Bricogne, G. *Acta crystallogr.* **A32**, 832–847 (1976).
- Weis, W. thesis, Harvard Univ. (1987).
- Knossow, M. *et al. Acta crystallogr.* **B42**, 627–632 (1986).
- Lesk, A. M. & Hardman, K. D. *Science* **216**, 539–540 (1982).
- Knossow, M., Daniels, R. S., Douglas, A. R., Skehel, J. J. & Wiley, D. C. *Nature* **311**, 678–680 (1984).

The intergalactic medium towards supernova 1987A

Paola Andreani* & Alfred Vidal-Madjar

Institut d'Astrophysique CNRS, 98bis bd Arago, 75014 Paris, France

The galactic halo and the intergalactic medium have been investigated by analysis of absorption features in the spectra of Seyfert galaxies, quasars, globular clusters, high-latitude stars and Magellanic Cloud supergiants¹. The halo gas is thus thought to be distributed in an expanded disk not more than 10 kpc thick. But radioastronomy observations have revealed other structures, the high-velocity clouds, which are probably associated with the galactic halo. They are seen particularly in the direction of the Magellanic Clouds, where the Magellanic Stream seems to merge²⁻⁵. The extraordinary brightness of supernova 1987A in the Large Magellanic Cloud (LMC) allowed us to obtain high-resolution spectra with high signal-to-noise ratios, sampling the galactic disk and halo, interstellar gas in the LMC, and the intervening medium. Thirty-five distinct components have been clearly identified, probably located between the Galaxy and the LMC, and showing low dust content.

Visible, ultraviolet and radio observations reveal many more features along the line of sight to the LMC than in other directions through the halo, and indicate also an unusual velocity structure⁵⁻⁸. In other directions only a few extreme velocities have been found; $v_{lsr} = 193 \text{ km s}^{-1}$ towards Fairall 9 (ref. 8) and $v_{lsr} = 241 \text{ km s}^{-1}$ towards the Seyfert galaxy NGC3783 (ref. 9), for example, where v_{lsr} denotes velocity with respect to our local standard of rest. These have been interpreted to indicate extragalactic material belonging to the Magellanic Stream. Our observations reveal the presence of a peculiar structure at heliocentric velocity $v_{hel} = 219 \text{ km s}^{-1}$, probably related to high-velocity gas inside the LMC. (To change v_{hel} to v_{lsr} subtract 15.5 km s^{-1} .)

The observations were gathered at the ESO CAT telescope in La Silla, Chile, during 24-26 February 1987, and have been described elsewhere^{10,11}. Here we report the detailed analysis of the Ca II K and H lines and Na I doublet spectra. Our most striking results are: (1) a great number of structures receding from us and covering the whole velocity range from the Sun up to the position of SN1987A were detected; (2) to fit the calcium line profile with gaussians, at least 35 cloud components are needed, (3) most of the 35 features found in calcium were not present in the sodium spectrum, especially at intermediate velocities; and (4) a particularly strong and double structure was detected ~ 216 and 219 km s^{-1} (heliocentric).

To separate the cloud components, a line-profile fitting method was used, taking into account the instrumental profile (the spectrograph has a triangle-shaped 3 km s^{-1} FWHM (full width-half-maximum) instrumental function; the estimated uncertainty on its width is $\sim 0.3 \text{ km s}^{-1}$), the shape of the continuum, and assuming that the line-absorption has a Voigt profile. Data are compared with convolution of these three terms. (Details on the iterative process can be found in ref. 12.) The best fit gives for each cloud the velocity, the column density and an estimate of the temperature using the b -value ($b = ((2kT/m) + \sigma_t^2)^{1/2}$, where σ_t is the line-of-sight turbulent velocity dispersion in the cloud).

Each line was fitted independently to find the minimum number of components required to reproduce the profile (the velocity, the b -value and the column density of each component are

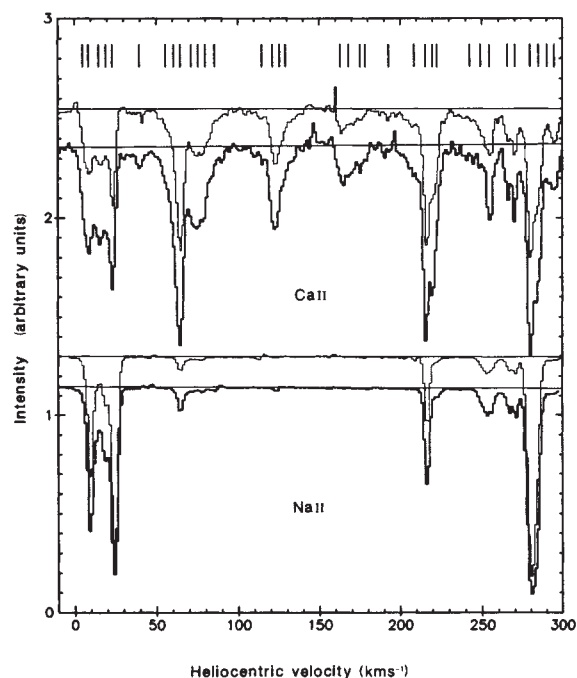


Fig. 1 The superposed Ca II K and H lines (above), and the Na I D1 and D2 lines (below) as functions of the heliocentric velocity. Thick marks at the top indicate the 35 components found by fitting the profile.

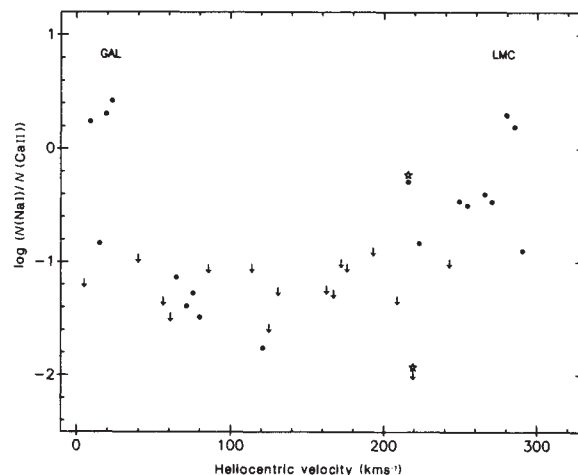


Fig. 2 The abundance ratio $N(\text{Na I})/N(\text{Ca II})$ as a function of the components heliocentric velocity. Upper limits are given where sodium was not detected. Features at $v = 0-25 \text{ km s}^{-1}$ belong to the galactic disk, at $v = 220-300 \text{ km s}^{-1}$ to the LMC interstellar gas; the location of the intermediate velocities structures is discussed in the text. Stars indicate the two 'anomalous' clouds discussed in the text.

always free parameters). The solution is not unique because a greater number of components with smaller b -values (the resolution achieved with this method is 1.5 km s^{-1}) fits the data as well, but this does not affect the evaluated total column densities. Weak features were rejected if they did not compare in both lines of the doublet. The results are listed in Table 1. The error in the derived column densities is $\sim 10\%$ (except for the weak components 21 and 22 and the blended ones 16 and 17, which are located in a particularly noisy region of the Ca II K-spectrum). The precision of this evaluation can be seen directly on the spectra (see Fig. 1). For instance, clouds 30 and 31 differ

* Present address: Dipartimento di Fisica, Università di Roma, La Sapienza, P.le Aldo Moro 2, 00185 Roma, Italy.

Table 1 Velocities, column densities and b -values from Ca II K and H lines and the Na I doublet

	Ca II v (km s ⁻¹)	K N (10 ¹⁰ cm ⁻²)	H N (10 ¹⁰ cm ⁻²)	Na I v (km s ⁻¹)	D2 N (10 ¹⁰ cm ⁻²)	D1 N (10 ¹⁰ cm ⁻²)	b^* (km s ⁻¹)
1	5.1	7.42	4.93	—	—	—	1.57
2	8.8	27.03	30.87	9.6	45.39	55.40	2.44
3	15.0	28.40	27.72	13.5	5.67†	2.53	2.84
4	19.2	7.47	8.59	17.9	16.23	16.43	1.70
5	23.0	36.33	43.55	23.8	90.00	121.70	2.02
6	39.8	3.42	4.06	—	—	—	2.16
7	56.2	9.00	8.87	—	—	—	3.15
8	60.8	13.74	10.94	—	—	—	2.03
9	64.8	66.02	71.24	64.8	5.09	4.90	2.23
10	71.4	18.53	18.90	70.9	0.73	0.79	2.61
11	75.7	15.79	17.53	75.8	0.66	1.09†	2.24
12	80.0	13.30	18.10	79.8	0.54	0.47	2.21
13	85.8	4.85	4.55	—	—	—	2.76
14	114.1	4.73	4.35	—	—	—	1.99
15	121.0	19.36	16.44	123.1	0.38†	0.23	1.92
16	125.0	13.28	17.41	—	—	—	2.02
17	131.0	6.42	8.19	—	—	—	2.52
18	162.5	7.24	6.91	—	—	—	2.39
19	167.0	8.98	6.40	—	—	—	3.30
20	172.0	4.25	3.96	—	—	—	2.52
21	176.0	4.82	4.21	—	—	—	2.24
22	192.8	2.46	4.02	—	—	—	2.21
23	208.6	8.27	9.20	—	—	—	2.51
24	215.7	49.26	45.80	216.4	23.50	25.10	1.82
25	219.0	35.48	44.70	219.0	<0.4	<0.4	2.85
26	222.8	9.42	9.42	223.9	1.28	1.47	2.26
27	242.7	4.10	4.11	—	—	—	1.42
28	249.2	10.30	9.61	249.7	2.81	4.05	2.34
29	254.4	19.86	19.55	253.9	5.95	6.50	1.87
30	265.5	9.28	11.37	266.1	3.87	4.32	1.76
31	270.3	12.72	16.79	270.6	4.72	5.33	1.52
32	279.7	72.14	67.99	279.7	112.92	166.60	2.07
33	284.9	36.13	46.31	283.0	63.14	65.78	2.13
34	290.3	7.91	6.36	291.0	0.89	0.90	2.18
35	295.0	9.40	12.94	—	—	—	1.69

* The b -value is deduced from the Ca II lines, and is assumed to be the same for the Na lines.

† Values affected by water vapour absorption.

by less than 20% in column density and are undoubtedly distinct. The velocity error bar is 0.5 km s⁻¹. The line width (with an error bar of ~0.5 km s⁻¹) shows that most of the components have temperatures lower than 10,000 K and that none is hotter than 20,000 K. This means we have sampled a relatively cool medium.

To infer the physical parameters of the clouds, one can use the sodium/calcium abundance ratio, which is mainly a function of the temperature, T , the electron number density, n_e , the radiation field and the calcium depletion factor, δ_{Ca} . Assuming that ionization is primarily due to photon absorption and that the Ca I, Ca IV and Na III abundances are negligible (as far as Ca I abundance in the supernova line of sight is concerned, this statement is confirmed by the measurements reported in ref. 13), it follows that the usual balance equations read:

$$\frac{N(\text{Na I})}{N(\text{H I})} = \frac{n_e \alpha_{Na} A_{Na}}{\Gamma_{Na}} \quad (1)$$

$$\frac{N(\text{Ca II})}{N(\text{H I})} = \frac{n_e \alpha_{Ca} A_{Ca}}{\delta_{Ca} (\alpha_{Ca} n_e + \Gamma_{Ca})} \quad (2)$$

and the resulting Na I Ca II ratio is

$$\begin{aligned} \frac{N(\text{Na I})}{N(\text{Ca II})} &= \frac{\alpha_{Na}}{\alpha_{Ca}} \frac{A_{Na}}{A_{Ca}} \frac{\Gamma_{Ca}}{\Gamma_{Na}} \left(\frac{\alpha_{Ca} n_e}{\Gamma_{Ca}} + 1 \right) \delta_{Ca} \\ &= \left[\frac{N(\text{Na I})/N(\text{H I})}{A_{Ca}} + \frac{\alpha_{Na}}{\alpha_{Ca}} \frac{A_{Na}}{A_{Ca}} \frac{\Gamma_{Ca}}{\Gamma_{Na}} \right] \delta_{Ca} \quad (3) \end{aligned}$$

where α is the temperature-dependent recombination coefficient,

A is the cosmic abundance and Γ is the photoionization coefficient.

Figure 2 shows the $N(\text{Na I})/N(\text{Ca II})$ values against the components' heliocentric velocities; upper limits are given where sodium was not detected. From values of ~2 (at v_{hel} near 0 and 300 km s⁻¹), typical of the standard interstellar medium¹⁴, the Na I/Ca II ratio drops to 10⁻². This confirms the reduction of the Na I/Ca II ratio in high-velocity material also observed along other lines of sight¹⁵⁻¹⁹. Furthermore, its value is equal to the theoretical lower limit (2×10^{-2}) (ref. 20), suggesting that no dust is present and that the Ca abundance is nearly solar in the detected structures.

To explain this, we estimated the elemental abundance by using the H I measurements available in the literature^{5,21-23}. In Table 2 we report the hydrogen column densities in directions less than 40 arc min from the supernova together with the corresponding ion or neutral abundance. By using the H I measurements and adopting the values $A_{Na} = 1.76 \times 10^{-6}$, $A_{Ca} = 2.14 \times 10^{-6}$ (ref. 24), $\Gamma_{Na} = 2.3 \times 10^{-11} \text{ s}^{-1}$, $\Gamma_{Ca} = 3.2 \times 10^{-12} \text{ s}^{-1}$ (ref. 25), it turns out that the $(N(\text{Na I})/N(\text{H I}))/A_{Ca}$ term, in equation (3), is negligible ($\sim 10^{-3}$) when compared to the next term, $\alpha_{Na}/\alpha_{Ca} A_{Na}/A_{Ca} \Gamma_{Ca}/\Gamma_{Na}$. The latter is nearly constant, ~70, because the α_{Na}/α_{Ca} ratio varies by at most a factor of 1.5 for T between 100 and 10,000 K (ref. 26) and the Γ_{Ca}/Γ_{Na} ratio is not a function of the distance from the Galaxy (unless its value is very different in the LMC from that in the Galaxy). We can rewrite equation (3) as

$$\delta_{Ca} = \frac{74 N(\text{Na I})/N(\text{Ca II})}{(3.4 \times 10^7 N(\text{Na I})/N(\text{H I}) + 1)} \quad (4)$$

The Ca depletion factor versus the heliocentric velocity thus changes exactly in the same way as $N(\text{Na I})/N(\text{Ca II})$ does. The observed behaviour can then be ascribed to the Ca depletion factor (see equation (4)) which is shown in Fig. 3.

Figure 2 also shows that a few of the components present anomalies. Components 1, 3 and 35 reflect the scatter in the gas abundance from one cloud to the other within the galactic disk and the LMC interstellar gas (they show lower depletion in Ca, by a factor of 10, than in the average interstellar medium). Components 24 and 25 shows respectively a high and low calcium depletion (see Fig. 3). By taking account of the higher $N(\text{Na I})/N(\text{H I})$ ratio of cloud 24 (see Table 2), which means

Table 2 Hydrogen column density and element abundances

Component	Velocity $v(\text{H I})$ (km s^{-1})	Density $N(\text{H I})$ (10^{20} cm^{-2})	Abundance ratio $N(\text{Ca II})/N(\text{H I})$ (10^{-10})	Abundance ratio $N(\text{Na I})/N(\text{H I})$ (10^{-10})
3	15.5	3.24	8.66	1.26
4-5	20.0	4.10	11.70	29.80
7	54.0	0.029	308.28	—
10-11	73.0	0.007	5,054.29	234.29
12-13	81.0	0.062	327.42	10.65
15-17	126.0	0.077	2,197.97	28.72
19-21	172.0	0.009	1,812.22	—
24-25	220.0	<0.007	>12,500.00	>3,500.00
27-28	247.0	7.94	1.77	0.45
28-29	253.0	0.776	38.23	12.45
31	270.0	8.440	1.75	0.60
31-32	275.0	17.900	4.74	8.09
32-33	282.0	6.92	16.08	29.51

that the term $[N(\text{Na I})/N(\text{H I})]/A_{\text{Ca}}$ in equation (3) is not negligible, the corresponding point in Fig. 3 can be lowered by at least a factor of 2 (see equation (4)).

On the other hand, cloud 25 does not show up in sodium, but has a strong calcium column density. Several checks have been done on the sodium spectrum: even by changing all the parameters, the profile-fitting never allows the presence of component 25 at 219 km s^{-1} . Its very small δ_{Ca} value (with respect to the others) confirms the peculiarity of this component. Although its peculiarity can be ascribed to its higher volume density, because, very probably, the electron number density is the changing parameter in the $\alpha_{\text{Ca}} n_e / \Gamma_{\text{Ca}}$ term of equation (3), the nature of the 25th component is clearly different. It could be identified with the anomalous metal-abundance feature detected by the International Ultraviolet Explorer²⁷ or with the anomalous electron density found by Magain¹³ in the 220 km s^{-1} cloud. Only high-resolution data together with profile fitting can distinguish the different origin of the two features, and the above conclusions do not depend on the inferred H I column density but only on the different Na I/Ca II ratio (see Fig. 2). Because H I was not detected at this velocity, the two points in Fig. 3 can only be lowered by increasing the $N(\text{Na I})/N(\text{H I})$ ratio.

Figure 3 seems to indicate a monotonic variation of the calcium depletion factor with the radial velocity. One can argue that a physical link exists among the clouds and that the mechanism responsible for the spread in the radial velocities could also control the dust content of each structure.

The coincidence of most of the components with the H I ones shows that they probably belong to the high-velocity structures already detected at 21 cm (refs 2-5, 21-23) and that most of the intermediate velocity features are part of the Magellanic Stream. McGee *et al.*⁵ deduced that, within the structures in the direction of the LMC, most of the observed elements (O I, Mg II, Al II, Si II, Fe II) show small or no depletion, while Ca II appears to be depleted in the strong components and less depleted in the

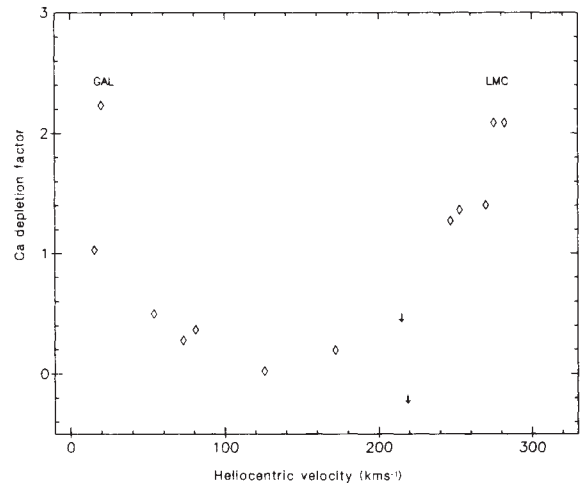


Fig. 3 Ca depletion factor, as defined in equation (4) against the heliocentric velocity of the H I features. Not all the 35 components are reported because hydrogen has not been detected at all velocities. Upper limits are shown for the 'anomalous' structures discussed in the text. The ordinate scale is logarithmic.

weaker ones. In addition, the lack of any calcium depletion in our weaker features suggests that these structures show normal metallicity and no dust grains.

In general, a nearly solar metallicity in the intermediate velocity components can be inferred from our data. If these structures are indeed associated with the Magellanic Stream, one can argue that this medium shows a nearly solar abundance in the detected elements. Our conclusion is in contrast with that drawn by West *et al.*⁹ for the same medium, who ascribe the low $N(\text{Ca II})/N(\text{H I})$ ratio to its intrinsic low metallicity or depletion to grains. Their structure, although related to the Magellanic Stream, is probably far away from the Magellanic Clouds.

We would like to thank R. Ferlet and L. Spitzer for very fruitful discussion and comments on this paper and M. Peron for her continuous help and assistance. P. A. is supported by an ESA fellowship.

Received 21 September 1987; accepted 31 March 1988.

- York, D. G. *A. Rev. Astr. Astrophys.* **20**, 221-248 (1982).
- Verschuur, G. L. *Astr. Astrophys.* **22**, 139-151 (1973).
- Giovannelli, R. & Haynes, M. P. *Mon. Not. R. astr. Soc.* **177**, 525-530 (1976).
- Mirabel, I. F., Cohen, R. J. & Davies, R. D. *Mon. Not. R. astr. Soc.* **186**, 433-451 (1979).
- McGee, R. X., Newton, L. M. & Morton, D. C. *Mon. Not. R. astr. Soc.* **205**, 1191-1205 (1983).
- Savage, B. D. & de Boer, K. S. *Astrophys. J.* **243**, 460-484 (1981).
- Songaila, A., Blades, J. C., Hu, E. M. & Cowie, L. L. *Astrophys. J.* **303**, 198-215 (1986).
- Songaila, A. *Astrophys. J.* **243**, L19-L22 (1981).
- West, K. A., Pettini, M., Penston, M. V., Blades, J. C. & Morton, D. C. *Mon. Not. R. astr. Soc.* **215**, 481-497 (1985).
- Andreani, P., Ferlet, R. & Vidal-Madjar, A. *Nature* **326**, 770-772 (1987).
- Vidal-Madjar, A. *et al. Astr. Astrophys.* **177**, L17-L20 (1987).
- Vidal-Madjar, A., Laurent, C., Bonnet, R. M. & York, D. G. *Astrophys. J.* **211**, 91-107 (1977).
- Magain, P. *Nature* **329**, 606-607 (1987).
- Hobbs, L. M. *Astrophys. J.* **191**, 381-393 (1974).
- Siluk, R. S. & Silk, J. *Astrophys. J.* **192**, 51-57 (1974).
- Blades, J. C. & Morton, D. C. *Mon. Not. R. astr. Soc.* **204**, 317-329 (1983).
- Morton, D. C. & Blades, J. C. *Mon. Not. R. astr. Soc.* **220**, 927-948 (1986).
- Cohen, J. G. & Meloy, D. A. *Astrophys. J.* **198**, 545-549 (1976).
- York, D. G. *et al. Astrophys. J.* **255**, 467-474 (1982).
- Hobbs, L. M. *Astrophys. J.* **265**, 817-823 (1983).
- Meaburn, J., Marston, A. P., McGee, R. X. & Newton, L. M. *Mon. Not. R. astr. Soc.* **225**, 591-606 (1987).
- Rohlfs, K., Kreitschmann, J., Siegmund, B. C. & Feitzinger, J. V. *Astr. Astrophys.* **137**, 343-357 (1984).
- McGee, R. X. & Newton, L. in *Proc. astr. Soc. Aust.* **6**, 358-385 (1986) in *Menzel Symp., Natl Bur. Stan. spec. Pub. 353* (ed. Gebbie, K. B.) 137 (US Government Print Office, Washington, 1971).
- Withbroe, G. L. in the *Menzel Symp.* (NBS Spec. Pub. 353) ed. K. B. Gebbie (Washington, US Government Print Office), p. 137 (1971).
- Herbig, G. H. *Zeit. für Astrophys.* **68**, 243-277 (1951).
- Seaton, M. *Mon. Not. R. astr. Soc.* **111**, 368-376 (1951).
- de Boer, K. S. *et al. Astr. Astrophys.* **177**, L37-L40 (1987).

Flux ratio of C₂ Swan bands in the innermost atmosphere of comet Halley

V. Vanysek*, B. Valniček† & J. Sudová†

* Astronomical Institute, University Erlangen-Nürnberg, Sternwartstrasse 7, 8600 Bamberg, FRG

† Astronomical Institute, Czechoslovak Academy of Science, 251 66 Ondřejov, Czechoslovakia

The spectrophotometry of comet Halley performed by the three-channel spectrometer (TKS) aboard the Vega 2 spacecraft provides spectrometric data with an unprecedentedly high angular and spatial resolution. Here we report observations of the flux-ratio variation of the C₂ Swan band sequence as it varies with distance from the comet nucleus. The flux in the C₂ ($\Delta v = +1$) band relative to C₂ ($\Delta v = 0$) increases toward the nucleus by at least a factor 1.6. Such an effect may be crucial for testing current theories of the resonance fluorescence radiation of the cometary C₂ molecule.

The TKS data used in this study were reduced from a set of digitized spectrograms obtained in the visible channel during the spacecraft's encounter with the comet on 9 March 1986. The instrument and some preliminary results have been described elsewhere¹⁻³. The field of view of the spectrometer ($2^\circ \times 1.5^\circ$) was scanned in 15 positions along 7 lines. During closest approach to the comet, scanning was performed only by a see-saw motion along line 4 (instrument mode B). For our study we chose data obtained at position 8 on line 4, which was situated almost exactly in the centre of the scanned field. The angular dimension of the area encompassed by the slit was $25' \times 0.5'$.

The optical axis of the spectrometer was inclined to the radius vector spacecraft-nucleus at an angle $\theta = 1.4^\circ$. Thus the instrumental line of sight had a minimum distance ρ_m from the nucleus (the impact parameter), which is defined with sufficient accuracy as $\rho_m = D(t) \sin \theta$, where $D(t)$ is the spacecraft distance from the nucleus at onboard time t .

On 9 March 1986 the instrument collected data between 5:33 and 7:20 UT. From this observational sequence we reduced a set of 34 spectra corresponding to the distance range $4.9 \times 10^5 \geq D(t) \geq 8.4 \times 10^3$ km and $1.2 \times 10^4 \geq \rho_m \geq 2.0 \times 10^2$ km.

The fluxes of the C₂ band were determined by integrating over the spectral features after correcting for the continuum. The measured flux $F(\lambda)$ was normalized to flux $F(\lambda_0)$ at $\lambda_0 = 488$ nm in a wavelength interval proportional to the width of the pixel or 0.61 nm, which is approximately 10% of the FWHM derived from the instrument response function. The reference wavelengths chosen for derivation of the underlying continuum by comparison with normalized intensities, taken from the Utrecht Atlas of the solar spectrum, were 440, 488 and 530 nm. Although there was some reddening of the comet's continuum owing to selective light scattering on dust particles, the flux distribution of the Sun and the cometary continuum in the studied spectral range differ no more than by a few per cent. At the long-wavelength end of the comet spectrum ($\lambda > 530$ nm) the reddening is more pronounced and the comet continuum is enhanced relative to the solar spectrum by a factor 1.10 to 1.14. It is compatible with results obtained by several other authors⁴⁻⁶.

Figure 1 shows the integrated fluxes corrected for the underlying continuum as the flux ratios of C₂ Swan bands, $F_R = F(\Delta v = +1)/F(\Delta v = 0)$ plotted against the impact parameter ρ_m . Each point in the range $\rho_m < 5 \times 10^3$ km represents a value averaged from two spectrograms. The dispersion of individual data points is $\sim 5\%$ and the undulating variation of the flux ratio with impact parameter is probably real.

Nevertheless, the contribution of the continuum increases greatly toward the nucleus, and uncertainties may have a system-

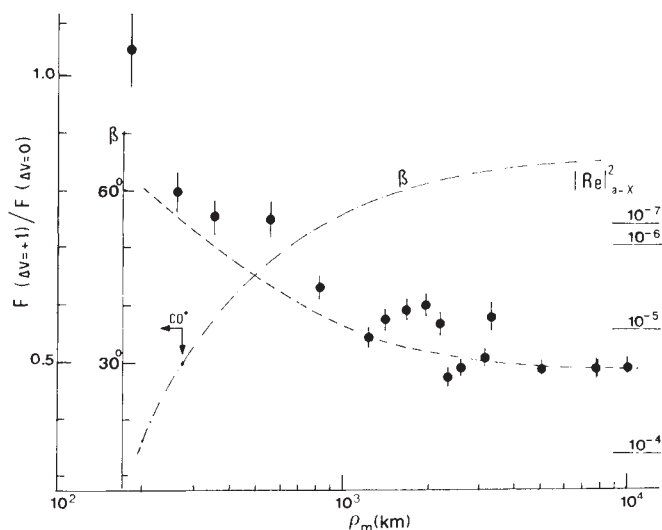


Fig. 1 The flux ratio of the C₂ ($\Delta v = +1$)/($\Delta v = 0$) Swan bands against impact parameter derived from Vega 2 TKS spectrograms. The dashed curve represents the 'smoothed' growth of the flux ratio toward the nucleus after a correction from the imperfectly subtracted continuum. The horizontal bars indicate the theoretical flux ratio computed by Krishna Swamy and O'Dell⁷ for various values of the electronic transition moment $|Rel|^2$. The dashed-dotted curve represents variation of the phase angle β (spacecraft-comet-Sun) with the impact parameter. The flux ratio corresponding to $\beta < 30^\circ$ may be affected by an unresolved emission of the CO⁺ band system of the comet's tail.

atic character due to imperfectly subtracted continuum flux. The response function (FWHM) of the visual channel was ~ 7 nm and the weak emission features cannot easily be discriminated against the continuum. Whereas the continuum intensity varies approximately with $1/\rho_m$, the emission profiles along the impact parameter at small ρ_m are almost flat. Therefore the 'wings' of bands blend into the underlying continuum as ρ_m decreases. A systematic variation of F_R with ρ_m may be a consequence of this effect. It may be assumed that owing to the spectral profile of C₂ bands the imperfectly reduced continuum varies exponentially with $1/\rho_m$, and an *ad hoc* correction factor, $f_c = 1 - k \exp(-\rho_m/\rho_0)$ can be adopted, where k is an empirical constant.

As F_R increases toward the nucleus, the $\Delta v = 0$ band should be more affected by the continuum contamination as $\Delta v = +1$. Thus the corrected flux ratio is $f_c F_R$. But the spectral profile of this band shows no significant distortion even at $\rho_m = 400$ km. The imperfectly subtracted continuum flux inferred from the dispersion of the individual data for impact parameter at this distance may be ~ 10 –15%. for scale distance unit 10^3 km the empirical constant k would be 0.3.

The dashed curve in Fig. 1 represents variation of smoothed and corrected data with the assumptions $k = 0.2$ and $\rho_0 = 10^3$ km. If this curve is considered as a lower limit of the corrected data, the growth of the flux ratio F_R towards the nucleus is real. This effect would be of a great importance for modelling the radiation process in C₂ molecules. As was shown in a recent review⁷, the fluorescence resonance radiation of C₂ in comets under an assumption of statistical equilibrium can be described by a very sophisticated model including single-triplet transitions. Nevertheless, those authors find that Swan sequence flux ratios are primarily determined by the rate at which transitions occur between the lowest electronic triplet state 'a' forming the Swan bands and X state, that is the lowest singlet level. (Only the prefix letter are here used for the notation of levels.)

This is a forbidden transition for which the electronic transition moment $|Re|^2$ is unknown and is assumed in the above-mentioned model as a free parameter. The horizontal bars in Fig. 1 indicate the flux ratio derived theoretically⁷ for $(\|Re\|_{(a-x)}^2) = 10^{-4}, 10^{-5}, 10^{-6}$ and 10^{-7} atomic units and interpolated for the heliocentric distance 0.83 AU. The mean value of F_R for $\rho_m > 2,500$ km is 0.53 ± 0.03 . The same results were obtained by Tegler and O'Dell⁸ for comet Halley at almost the same heliocentric distance but from ground-based observations, which provide only averaged values for a relatively large area of the coma (that is, $\sim 10^9$ km² in this case). But in the innermost dense coma, where $\rho_m < 2,000$ km, the flux ratio according our results was ~ 0.75 , which is identical with the theoretical value if only triplet transitions in the Swan bands are involved and no leaking through any other states occurs.

The flux ratio derived from our data for the $(\Delta v = -1)/(\Delta v = 0)$ bands is ~ 0.5 and shows no systematic variation with ρ_m if the reddening of the comet's continuum is subtracted. Unfortunately the scattering of the individual values is considerably larger than that of F_R data and therefore we do not discuss them here in detail. Moreover, as suggested by other authors², the C_2 ($\Delta v = -1$) band could be contaminated in the inner coma by the NH_2 (0, 11, 0) emission as this molecule appears strongly concentrated toward the nucleus. The similar molecular contamination of $\Delta v = 0$ and $\Delta v = +1$ bands seems to be less significant. Nevertheless, if the line of sight is almost parallel to the cometary ion tail axis, the emission of CO^+ could be a source of an additional flux in the spectral range of the C_2 bands. The CO^+ (2-1) band at 470–473 nm is situated almost in the middle of the C_2 ($\Delta v = +1$) and CO^+ (1-1) emission at 507 nm on the short-wavelength side of the C_2 ($\Delta v = 0$). Although the presence of CO^+ bands in our spectra cannot be definitely confirmed, we believe that in the data shown here for the Swan bands' ratio only the highest value of the relative flux (~ 1.05), corresponding to the minimal value of the impact parameter, is likely to be affected by the emission of ionized carbon monoxide from the comet's tail. In this case, however, the phase angle (spacecraft-comet-Sun) was only 15° and the column density of CO^+ along the line of sight may have been considerably higher than that associated with other data for which the phase angle was 30 – 66° (see Fig. 1).

From the behaviour of the C_2 ($\Delta v = +1$) and ($\Delta v = 0$) Swan bands one can conclude that in the dense innermost cometary atmosphere even a sophisticated model based on the assumption of statistical equilibrium in the cometary C_2 is inconsistent with the observations. Because dissociation and chemical processes occur within the inner coma, the population distribution is dominated more by formation conditions than by radiative processes.

The electronic structure of diatomic carbon is characterized by strongly forbidden singlet-triplet transitions. The possibility cannot be excluded that in the dense region near the nucleus of comet Halley other processes than resonance fluorescence, induced by collisions, play a role in the statistical equilibrium of C_2 . One possible explanation is that such a mechanism may eliminate the singlet-triplet transition in the lowest state.

Large-amplitude photometric variations of Nereid

Martha W. Schaefer* & Bradley E. Schaefer†

* Laboratory for Terrestrial Physics, Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

† Laboratory for High Energy Astrophysics, Goddard Space Flight Center, Greenbelt, Maryland 20771, USA

Nereid, one of the two known satellites of Neptune, was discovered in 1949 by Kuiper¹. It has an extremely eccentric orbit ($e \approx 0.75$), such that its distance from Neptune varies from ~ 1.4 to 9.7 million km, with a period of ~ 360 days. No photometric data exist for Nereid² other than the rough estimate of photographic magnitude ($m_p = 19.5$) determined by Kuiper¹. In June 1987 we collected photometric (UBVRI) and astrometric data on Nereid using the 0.9-m telescope and CCD camera at the Cerro Tololo Inter-American Observatory (CTIO). We obtained four astrometric positions and UBVRI reflectances spanning 8 days. From these data, an unusual reflectance spectrum for Nereid was determined. Variability of Nereid was observed in all bandpasses, with an amplitude of greater than 1.5 mag, and a probable period of between 8 and 24 h.

The most recent orbit determination for Nereid³ is based on a set of only 50 positions collected between 1949 and 1981 by several observers. These osculating elements were estimated³ to become invalid to a precision of 1 arc s in at least one coordinate by 1989. In August of 1989, Voyager 2 will pass within 35,000 km of Neptune and within 40,000 km of Triton. It will then have the opportunity to look at Nereid as well, which will be only about a month past periapsis. To improve pointing accuracy, an accurate orbit for Nereid is necessary. An orbit error of 1 arc s at the distance of Earth translates to an error in position of the satellite of over 20,000 km, and an error in Voyager pointing of the order of half a degree.

Nereid was observed using the 0.9-m telescope at CTIO, 18–26 June 1987. A CCD camera at $f/13.5$, with resolution of ~ 0.5 arc s per pixel and a usable field of 396×396 pixels was used in data collection. Nereid was found without difficulty by comparing frames separated by an hour in time. Three frames showed Nereid to be unresolved from relatively bright stars, and were discarded. The possible presence of confusing background stars was checked for by comparison of frames with Nereid in different positions against the background stars. On three frames, Nereid was unresolved from a star of comparable magnitude. The brightness of the confusing star was measured from other frames, then subtracted to give Nereid's brightness.

The six primary astrometric standard stars used were measured by Klemola⁴ and are expected to be in error by 0.2–0.3 arc s (or less⁵) on average in each coordinate. The four astrometric observations were made when the six Klemola primary standard stars were centred in the field of view, and when Nereid was within 70 arc s of all standard stars. Nereid was 100 arc s away from Neptune, so that glare was not a source

Received 2 December 1987; accepted 28 March 1988.

1. Krasnopolsky, V. A. *et al.* *Nature* **321**, 269–271 (1986).
2. Moreels, G. *et al.* *Nature* **321**, 271–273 (1986).
3. Moreels, G. *et al.* in *Proc. 20th ESLAB Symp. Explor. Comet Halley* (ESA-SP-250) Vol. 1, 451–453 (1986).
4. Jewitt, D. C. & Meech, K. J. in *Proc. 20th ESLAB Symp. Explor. Comet Halley* (ESA-SP-250) Vol. 2, 47–51 (1986).
5. Parisot, J. P. *et al.* in *Proc. Symp. Diversity and Similarity of Comets* (ESA-SP-278) 255–262 (1987).
6. Giese R. H. *et al.* in *Proc. 20th ESLAB Symp. Explor. Comet Halley* (ESA-SP-250) Vol. 2, 53–57 (1986).
7. Krishna Swamy, K. S. & O'Dell, C. R. *Astrophys. J.* **317**, 543–550 (1987).
8. Tegler, S. C. & O'Dell, C. R. *Astrophys. J.* **317**, 987–991 (1987).

Table 1 Astrometry

UT of mid-exposure 23 June 1987	Right ascension (obs) 18 h 26 m	Declination (obs) –22° 14'	ΔRA (obs-calc)	ΔDec (obs-calc)
4:28	52.755 ± 0.040 s	$26.26'' \pm 0.18''$	0.204 s	0.15"
4:46	52.659 ± 0.040 s	$26.36'' \pm 0.18''$	0.191 s	0.19"
5:37	52.425 ± 0.041 s	$26.45'' \pm 0.19''$	0.202 s	0.09"
5:54	52.339 ± 0.040 s	$26.54'' \pm 0.19''$	0.214 s	0.11"

All positions are referred to the epoch of 1950.0.

of error in the astrometry. These observations were made with air masses ranging from 1.01 to 1.03, so no refractive corrections were necessary. Short exposures (90 s) were taken to determine the positions of the secondary standard stars relative to the primary standards because, on the long (15-min) exposures necessary to observe Nereid the Klemola stars become saturated. The position of Nereid was determined both relative to the secondary standard stars and relative to the saturated primary standard stars directly. No significant difference was found in the positions determined using the two methods. The intensities and the centres of the star images on the CCD pictures were determined using the program DAOPHOT⁶. Nereid's position and brightness was determined as part of a simultaneous fit of all nearby stars to a point spread function evaluated for several moderately bright and uncrowded field stars. The identification of nearby stars was made with frames taken when Nereid was not in the field, and is complete to within roughly a magnitude of Nereid's brightness. The four astrometric positions determined are shown in Table 1. These positions differ from positions predicted by the JPL Neptune ephemeris DE-118, with Nereid offsets calculated using the orbit of Rose⁷, by 3 arcs in right ascension and 0.2 arcs in declination.

Photometric standard stars⁸ were observed with U, B, V, R and I filters and a wide range of air masses, to make accurate air mass and colour corrections. Photometric observations were made of Nereid at many times, with exposures at different bandpasses generally following one another closely in time (Table 2). Colour equations suitable for the filters used were derived from measurements of the standard stars. The brightness of Nereid is seen to vary similarly in all colours (Fig. 1). The range in mean opposition magnitude ($V(1, 0)$) is from 3.28 to 4.86 (with a mean of 4.06) giving an observed amplitude of >1.5 mag. The total amplitude of the variation may well be larger as we may not have observations from both a maximum and a minimum. The variations are likely to be due to rotation effects, with a rotation period that seems to be between 8 and 24 h. Attempts to fit a period to the data by means of a Fourier transform were ambiguous, due to the sparseness of the data.

The observed variability in reflectivity of Nereid could be caused either by albedo differences on its surface (similar to those on Iapetus) or by a non-spherical shape. The roughly estimated period for this variability is consistent with either of these causes.

If the variability is caused solely by a non-spherical shape, a ratio of areas (maximum cross-section/minimum cross-section) of over four is implied. The most conservative case, assuming a prolate spheroidal shape for Nereid, with radii $r_1 > r_2 = r_3$, and Nereid rotating around the axis of r_3 (perpendicular to the line between Earth and Nereid), an albedo of 1.0 and an amplitude of variability of 1.5 mag, implies $r_1 \approx 338$ km, and $r_2 = r_3 \approx 79$ km. A lesser albedo would produce a correspondingly larger size. But satellites the size of Mimas ($r = 195$ km) or larger are

Table 2 Photometry

Time*	Bandpass	Nereid magnitude	$V(1.0)^\dagger$
4.00	V	19.27 ± 0.06	4.54
4.32	B	19.92 ± 0.04	
4.60	R	18.92 ± 0.08	
4.88	U	20.23 ± 0.07	
5.17	V	$18.98 \pm 0.06^\ddagger$	4.25
5.43	I	$18.50 \pm 0.07^\S$	
5.77	U	19.71 ± 0.06	
6.03	U	19.87 ± 0.07	
6.32	B	19.67 ± 0.08	
6.60	R	18.49 ± 0.08	
6.87	I	18.37 ± 0.20	
124.47	V	18.42 ± 0.39	3.69
124.77	V	19.06 ± 0.21	4.33
125.62	V	18.63 ± 0.15	3.90
125.90	V	19.10 ± 0.15	4.37
126.72	V	$19.59 \pm 0.23^\parallel$	4.86
194.02	V	18.01 ± 0.02	3.28
194.28	V	18.01 ± 0.03	3.28

* Middle of exposure in hours after 0 h UT, 18 June 1987.

[†] At 0 h UT, June 18, 1987, Neptune-Earth distance was 29.23 AU, Neptune-Sun distance was 30.23 AU, and solar phase angle at Neptune was 0.36° .

[‡] Unresolved with a star of $V = 20.07$.

[§] Unresolved with a star of $I = 18.38$.

^{||} Unresolved with a star of $V = 19.32$.

Table 3 Reflectances

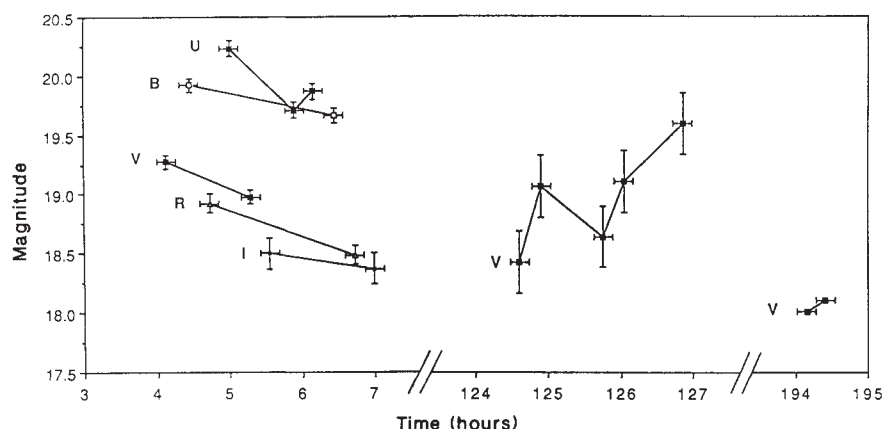
Colour	Sun	Nereid	Nereid reflectance ratio
U-V	0.84	1.07 ± 0.13	$U/V = 0.81 \pm 0.09$
B-V	0.64	0.93 ± 0.12	$B/V = 0.77 \pm 0.08$
V-R	0.31	0.18 ± 0.10	$R/V = 0.90 \pm 0.08$
V-I	0.73	0.42 ± 0.09	$I/V = 0.75 \pm 0.07$

all closely spherical⁹. This would seem to imply that the observed brightness contrasts are not due to shape alone.

If the variability is caused solely by albedo differences, an albedo ratio ('bright side'/'dark side') of at least four is implied (the albedo ratio of Iapetus is >10). This ratio could be produced by a spherical body of radius 163 km with an albedo on one side of 1.0 and on the other side of <0.24 .

The colours of Nereid were estimated by interpolation from the magnitudes at closely separated times, assuming a similar slope for the variation in all colours, and compared to the Sun's colours¹⁰ to obtain reflectances, shown in Table 3. In Fig. 2, the UVB spectrum of Nereid is compared to that of the classes of asteroids and that of various outer planet moons. The position of Nereid on the U-B vs B-V diagram is unlike that of any

Fig. 1 Light curve of Nereid in UBVR_I, showing large-amplitude variations.



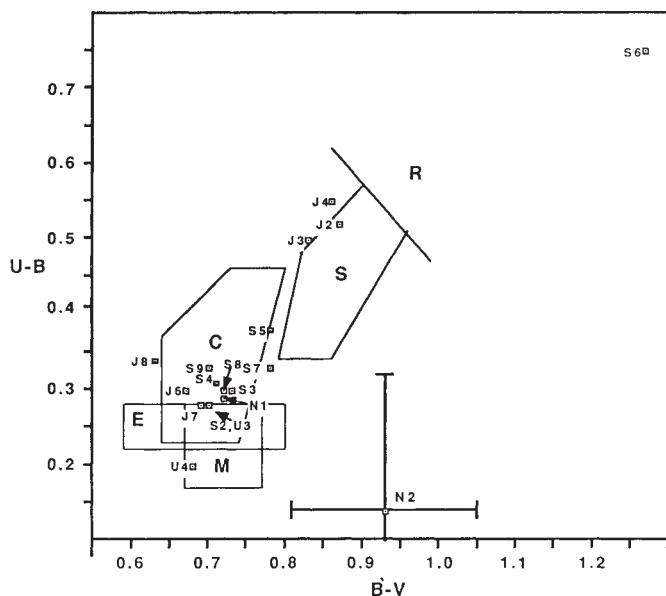


Fig. 2 Colour-colour diagram showing asteroids (after Bowell *et al.*¹⁷), outer planet moons¹⁸ and Nereid (this work). Nereid is quite different in colour from other small bodies in the solar system.

other satellite or asteroid. The red part of Nereid's spectrum is qualitatively different from that of these bodies as well.

One explanation for the unusual orbit of Nereid is that it is a captured asteroid. If this were true, however, the asteroid would have to be unusual in many ways: first, even for our smallest allowable radius, the captured asteroid would have to be one of the largest in existence; second, it would have to have an extremely unusual reflectance curve; and third, the amplitude of the light curve is larger than almost any asteroid (only a few small asteroids show comparable amplitude variations¹¹). Clearly, it is extremely unlikely that Nereid is a captured asteroid.

The remaining possibility is that Nereid accreted around some planet (presumably Neptune) as a satellite. We consider it unlikely that a satellite could accrete in an orbit of such high eccentricity and inclination¹². Therefore, Nereid must have suffered some drastic change in its orbit in the past. Large orbital changes in the Neptunian system have already been proposed¹³⁻¹⁶ to explain various peculiarities.

In summary, we have made four astrometric measurements of Nereid and observed a large amplitude variation likely due to an albedo difference between hemispheres of greater than a factor of four. Our observations provide strong evidence that Nereid is not a captured asteroid and hence indicate that Nereid has undergone a radical orbit change in the past.

We thank Arnold Klemola for providing us with astrometric standard star positions, Doug Mink for calculating Neptune and Nereid positions, and CTIO for providing observing time and assistance. Helpful discussions were had with Lucy McFadden, Dan Pascu, Rick Binzel and Guy Consolmagno. This work was funded by NASA.

Received 23 December 1987; accepted 29 March 1988.

1. Kuiper, G. P. *Publs Astr. Soc. Pacif.* **61**, 175-176 (1949).
2. Cruikshank, D. P. *NASA Conf. 2330: Uranus and Neptune* (ed. Bergstrahl, J. T.) 425-436 (NASA, Washington, 1984).
3. Veillet, C. *Astr. Astrophys.* **112**, 277-280 (1982).
4. Mink, D. J. & Klemola, A. *Astr. J.* **90**, 1894-1899 (1985).
5. Nicholson, P. D., McLeod, B. A., Gilmore, G., Buie, M. W. & Matthews, K. *Astr. J.* **95**, 562-575 (1988).
6. Stetson, P. B. *Publs Astr. Soc. Pacif.* **99**, 191-222 (1987).
7. Rose, L. E. *Astr. J.* **79**, 489-490 (1974).
8. Graham, J. A. *Astr. J.* **94**, 244-265 (1982).
9. Thomas, P., Veverka, J. & Dermott, S. in *Satellites* (ed. Burns, J. A. & Matthews, M. S.) 802-835 (University of Arizona Press, Tucson, 1980).

10. Degewij, J., Cruikshank, D. P. & Hartmann, W. K. *Icarus* **44**, 541-547 (1980).
11. Tedesco, E. F. in *Asteroids* (ed. Gehrels, T.) 1098-1107 (University of Arizona Press, Tucson, 1979).
12. Stevenson, D. J., Harris, A. W. & Lunine, J. I. in *Satellites* (ed. Burns, J. A. & Matthews, M. S.) 39-88 (University of Arizona Press, Tucson, 1986).
13. Harrington, R. S. & Van Flandern, T. C. *Icarus* **39**, 131-136 (1979).
14. Farinella, P., Milani, A., Nobili, A. M. & Valsecchi, G. B. *Icarus* **44**, 810-812 (1980).
15. Dormand, J. R. & Woolfson, M. M. *Mon. Not. R. astr. Soc.* **193**, 171-174 (1980).
16. McKinnon, W. B. *Nature* **311**, 355-358 (1984).
17. Bowell, E., Chapman, C. R., Gradie, J. C., Morrison, D. & Zellner, B. *Icarus* **35**, 313-335 (1978).
18. *The Astronomical Almanac for the Year 1988* (US Government Printing Office, Washington, 1987).

Extreme ionospheric effects in the presence of high electric fields

S. Buchert* & C. La Hoz†

* Max-Planck-Institut für extraterrestrische Physik,
D-8046 Garching bei München, FRG

† EISCAT, N-9027 Ramfjordbotn, Norway

In the presence of high-convection electric fields, theoretical considerations and computer models predict, among other effects, large increases in the ion temperature, the molecular ions and horizontal neutral winds in the high-latitude F region, but experimental evidence of these effects has been difficult to obtain. Here we report on measurements of very high perpendicular convection velocities of $\sim 3.8 \text{ km s}^{-1}$, corresponding to electric fields of 190 mV m^{-1} , that were made in the auroral F region of the ionosphere with the tristatic incoherent scatter radar at EISCAT in northern Scandinavia under magnetically perturbed conditions. These high velocities were associated with very high ion temperatures of 12,000 K without being accompanied by commensurate changes in the electron temperature. The large increase of the ion temperature can be accounted for quantitatively by Joule heating. The constraint of energy conservation during this event implies a substantial increase of molecular ions in the F layer, attaining values of more than 70% at 250 km altitude. Similarly, momentum conservation allows us to estimate neutral winds of up to 600 m s^{-1} mainly in the zonal direction.

The EISCAT incoherent scatter radar has been described by Rishbeth and Williams¹. In our experiment, the intersection altitude of the three beams was alternated between 165 and 250 km, with more frequent measurements at 250 km. The remote stations made measurements at only these altitudes. The transmitter station at Tromsø measured altitude profiles of the plasma parameters between about 150 and 350 km. The angle between the scattering vector and the Earth's magnetic field (aspect angle) varied between 2 and 70 degrees.

Figure 1 shows the perpendicular electric field and the corresponding ion temperature at an altitude of 250 km. During the time period between 0308 and 0402 UT (universal time), all the plasma parameters were strongly modulated by the passage of auroral forms known as Ω -bands². The enhancements of the electric field and the ion temperature shown in Fig. 1 occurred during the traversal of one of the dark areas of the Ω -bands where no precipitation occurs, whereas the lower values occurred coincidentally with a period of strong precipitation². Thus, the time variation of the parameters shown in Fig. 1 was determined by the velocity of the Ω -band traversal. The altitude profiles of the ion temperature measured at Tromsø repeatedly attained maxima at altitudes other than 250 km and reached values of up to 12,000 K. The largest ion temperatures to date have been recently reported as reaching about 8,000 K (ref. 3).

St-Maurice and Schunk⁴ have postulated the occurrence of non-thermal ion velocity distributions when the electric field-induced ion drift is larger than the neutral thermal speed and the ions are magnetized, typically taking place at a few tens of mV m^{-1} and above 150 km. Experimental evidence has recently

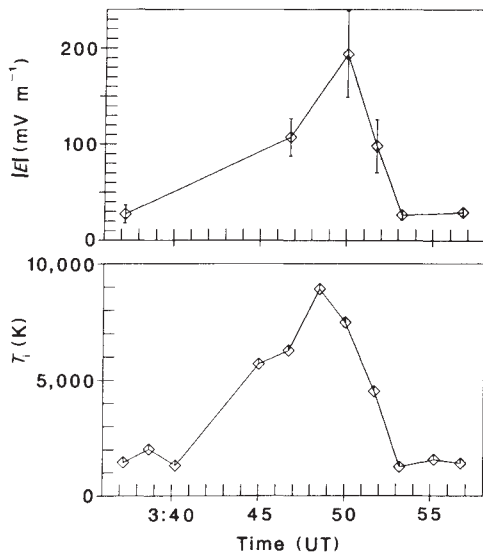


Fig. 1 Electric field strength perpendicular to the magnetic field (upper panel) and ion temperature (lower panel) as a function of universal time (UT) at an altitude of 250 km. The measurements were made on 21 April 1985 by the UHF EISCAT tristatic radar. The ion temperature, as well as other plasma parameters, is derived from the spectral shape of the scattered radar signal, and the electric field by combining the line-of-sight plasma velocities measured by the three EISCAT stations¹.

accumulated from incoherent scatter measurements made by EISCAT (refs 5 and 6). Thus we expect the ion velocity distribution to depart significantly from a thermal distribution. The directly measured scatter spectrum depends in a complicated manner on the three-dimensional ion velocity distribution. Consequently, our estimates of the plasma parameters are subject to systematic errors, as well as to ambiguities of interpretation. Raman *et al.*⁷ have found that the Maxwellian analysis of non-thermal spectra overestimates the ion temperature. Our results indicate that, as the ion temperature increases beyond the values considered by Raman *et al.*⁷, the non-thermal spectrum becomes progressively similar to the corresponding Maxwellian spectrum. Moreover, at values of the ion temperature as high as 11,000 K, the non-thermal analysis produces ion temperatures consistently larger than the Maxwellian analysis by about 1,500 K. Inspection of the dependence on the aspect angle reveals that the line-of-sight ion temperatures are overestimated at angles greater than 54° and underestimated for smaller angles⁷. Thus, for these high temperatures and the aspect angle at which they have been observed (52°), the systematic error is in a direction to underestimate the ion temperature. These calculations have been carried out using a value for the electron temperature equal to the background value, that is, an average at times before and after the heating events. This assumption is supported by the observation that the Maxwellian analysis during the heating event produced not very different values although sometimes somewhat larger. In summary, when the ion temperature is sufficiently high, the Maxwellian and non-thermal spectra are very similar.

Owing to the rapid convective motion of the ions through the more dense neutral background, exchange of energy and momentum takes place. The ions, due to ohmic (or Joule) heating, increase their temperature in a few seconds, whereas the increase of the convective velocity of the neutrals due to ion drag occurs over a few hours. The events we are considering last from a few minutes to a few tens of minutes. Under these conditions, ohmic heating is balanced by ion-neutral cooling. Thus under very general conditions, when the collisions are

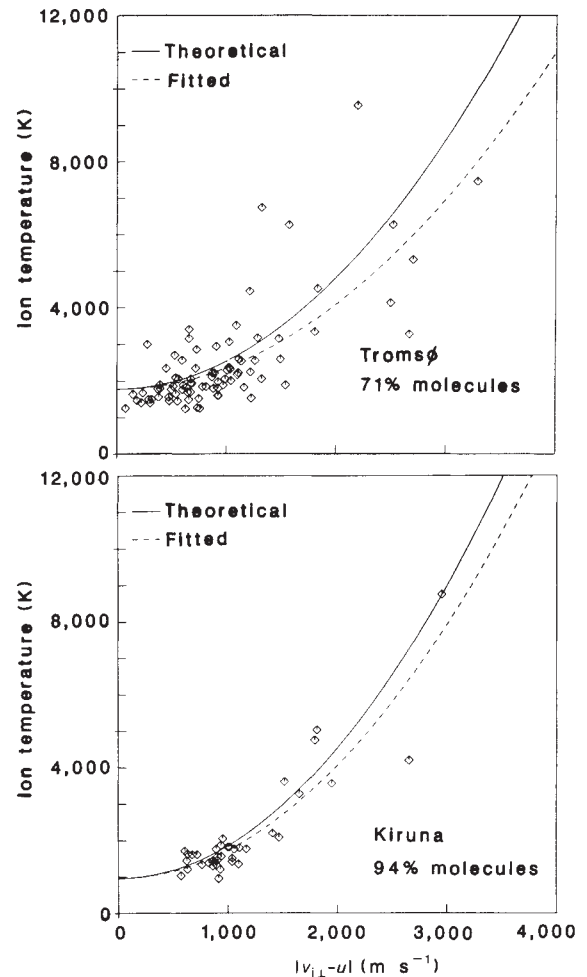


Fig. 2 Ion temperatures for altitudes 250 km (upper panel) and 165 km (lower panel) as a function of the velocity difference between the ion drift and the neutral velocity, $|v_{\perp} - u|$, in a plane perpendicular to the magnetic field. The diamonds are the measured values between 0000–0400 UT on 21 April 1985, the dashed line is a fitted parabola, and the continuous line represents equation (1), using the neutral temperature from the fit.

elastic, and neglecting heat conduction and energy losses to the electrons, the temperatures in the steady state are

$$k_B T_i = k_B T_n + \frac{m_n}{3} |v_{\perp} - u|^2 \quad (1)$$

where T_i and T_n are the temperatures of the ions and neutrals, respectively, u is the neutral velocity, k_B is the Boltzmann constant, m_n the average mass of the neutral gas, $|v_{\perp}| = |E_{\perp}|/B$ the ion drift velocity, $|E_{\perp}|$ the component of the electric field in the plane perpendicular to the magnetic field, and B the geomagnetic field strength. To test our measurements against equation (1), it is necessary to have an estimate of the neutral velocity. We have integrated an approximate momentum equation, based only on ion drag and neutral inertia, $\partial_t u = (\rho_i/\rho_n) \nu_{in} (v_{\perp} - u)$, where ρ is mass density and ν_{in} is the ion-neutral collision frequency⁸. The integration was started at 2130 UT to account for earlier acceleration of the neutrals, as electric fields of some tens of mV m⁻¹ were present since that early time. The results indicate substantial acceleration of the neutral atmosphere due to ion drag. By the end of the Ps 6 event at ~0400 UT, the neutral velocity at 250 km altitude attained values of ~600 m s⁻¹ in the eastward direction.

The high convection velocities and the increased heating of the plasma due to high electric fields have the effect of accelerat-

ing the rates of charge exchange reactions which produce the molecular ions NO^+ and O_2^+ , thus increasing the concentration of these constituents⁹. Therefore we have calculated T_i (and T_e) from the measured incoherent scatter auto-correlation functions by assuming various percentages of molecular ions, and compared the outcome with equation (1). The results are shown in Fig. 2 for the two altitudes at which vector velocities were measured, 165 and 250 km. The plotted points (diamonds) correspond to measurements using the values of u previously calculated. The dashed line is a parabola fitted to the measured data and the continuous line represents equation (1), taking the neutral temperature from the fit. We have found that the data at 250 km altitude are more consistent with equation (1) when the molecular ion concentration is greater than 70%. By contrast, measurements and accepted models at low and middle latitudes during quiet conditions usually give ~25% of molecular ions at 250 km. On the other hand, there is very good agreement between equation (1) and the data at 165 km altitude for molecular ion concentration of 94% or more.

The larger variation of the data points at 250 km compared to those at 165 km can be explained by the fact that the data cover an interval a few hours long, during which actual conditions were changing. These conditions affect the lower altitude much less because here the molecular ion composition is very high, even for small electric fields. The larger errors in the measurement of the incoherent scatter auto-correlation function at the higher altitude must make a contribution to the increased dispersion of the data points, since the signal-to-noise ratio decreases rapidly with altitude.

Here we have shown that convection events exist in the F region of the ionosphere, with $|\mathbf{E}_\perp|$ up to 190 mV m^{-1} being accompanied by rapid heating of the ions up to 12,000 K.

Additional calculations show that the molecular ion composition and the zonal neutral wind increase dramatically. These results represent a concrete illustration of the controlling role that geomagnetic electrodynamics exert on the thermosphere. This conclusion agrees with recent results from cooperative measurements and thermospheric computer models that include ionospheric-thermospheric coupling (see ref. 10 for summary and further references). Events of this type are possibly connected to recent observations by the AMPTE spacecraft of the existence of molecular ions in the plasma sheet made at about the same time as our measurements (E. Möbius, personal communication) and previous measurements¹¹, because an intermediate mechanism is necessary to transport or create molecular ions at the altitudes where ion acceleration mechanisms, such as ion conics and ion beams, are known to be effective.

We thank G. Haerendel for his continuing support and useful comments. We thank the director and staff of EISCAT for valuable assistance. EISCAT is an international cooperation supported by the French CNRS, West German MPG, UK SERC, Norwegian NAVF, Swedish NFR, and Finnish SA.

Received 1 February; accepted 19 April 1988.

1. Rishbeth, H. & Williams, P. J. S. *Q. Jl. R. astr. Soc.* **26**, 478–512 (1985).
2. Buchert, S., Baumjohann, W., Haerendel, G., La Hoz, C. & Lühr, H. *J. atmos. terr. Phys.* (in the press).
3. Kofman, W. & Lathuillière, C. *Geophys. Res. Lett.* **14**, 1158–1161 (1987).
4. St-Maurice, J.-P. & Schunk, R. W. *Rev. Geophys. Space Phys.* **17**, 99–134 (1979).
5. Lockwood, M. *et al. Geophys. Res. Lett.* **14**, 111–114 (1987).
6. Winsor, K. L. *et al. Geophys. Res. Lett.* **14**, 957–960 (1987).
7. Raman, R. S. V. *et al. J. geophys. Res.* **86**, 4751–4762 (1981).
8. Banks, P. M. & Kockarts, G. *Aeronomy, Part A* (Academic Press, New York, 1973).
9. Schunk, R. W. *et al. J. geophys. Res.* **81**, 3271–3282 (1976).
10. Rycroft, M. J. *Nature* **326**, 747–748 (1987).
11. Craven, P. D. *et al. J. geophys. Res.* **90**, 7599–7605 (1985).

Diamonds in detonation soot

N. Roy Greiner*, D. S. Phillips*, J. D. Johnson*
& Fred Volk†

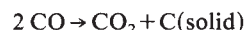
*Los Alamos National Laboratory, Los Alamos,
New Mexico 87545, USA

†Fraunhofer-Institut für Treib- und Explosivstoffe,
D-7507 Pfinztal/Berghausen, FRG

The physical structure and chemical bonding of the carbon in solid detonation products (soot) are largely unknown. It is well established that diamond can be manufactured by the application of explosive shocks to graphite loaded into the explosive, or in a fixture external to the explosive¹. Here we report the formation of diamonds as a chemical product of the detonation process itself. The diamonds we observe are 4–7 nm in diameter and make up 25 wt% of the soot; in size and infrared spectrum they resemble diamonds similarly isolated from meteorites².

The chemical nature of detonation soot has been a subject of interest for some time^{3–5}, and the formation of diamonds in other dynamic systems has received considerable attention recently^{2,6}. The chemical process in a detonating system is somewhat complex. The products form at high density, pressure, and temperature and then expand and cool isentropically (Table 1). Detonation of CHNO explosive compositions underbalanced relative to CO ($\text{O/C} < 1$, after complete oxidation of all H present to H_2O), such as those reported here, nearly always yield some solid carbonaceous residue (soot) as a recoverable product⁵. Calculations using the Becker-Kistiakowsky-Wilson (BKW) equation of state predict that the detonation of explosives initially near crystal density yield a product composition mostly of N_2 , H_2O , CO_2 and solid carbon at the Chapman-

Jouguet (CJ) point of the detonation wave, because equilibrium favours the reaction



at the very high CJ pressure ~300 kbar⁷. The model further predicts that some of the solid carbon and CO_2 formed at the CJ point will be consumed by the reverse reaction, and that some of the H_2O will react with CO in the water-gas reaction as the products equilibrate during expansion along the release isentrope⁷. These equilibria are believed to quench at temperatures somewhat below 2,000 K, freezing the composition of the recoverable products. The net effect is that nearly all but the most highly oxygen-balanced explosives will make carbon in a high-density detonation, but that some carbon predicted at the CJ point will be consumed as the products expand. Typically, the solid product is found to contain some nitrogen and hydrogen as well as carbon⁵. Note that the soot is formed in a hot high-pressure fluid of N_2 , CO_2 and H_2O .

The dynamics of detonation product expansion, and therefore the work available from the explosion, are dependent on the kinetics and thermodynamics of the formation and on the alteration of solid carbon as the expansion proceeds^{3,4,7,8}. Furthermore, the CJ state is also sensitively dependent on the densities of any solid products present. At the CJ state, the density of the products is about 1.35 times the density of the original solid explosive, so the pressure and temperature are strongly dependent on the volume taken up by each product component. Previous studies have indicated that changes from the graphite to diamond phase influences the detonation velocity and final state pressures by 5–10% and 10–20% respectively^{3,4}. The break in slope of the detonation velocity versus initial density of trinitrotoluene (TNT) at about 1.55 g cm^{-3} has been attributed to changes in the carbon equation of state^{7,9}.

This strong dependence of detonation behaviour on the character of carbon in the high-temperature (~3,000 K), high-

Table 1 Sample data

Sample number	27	60	63
Composition	40% TNT/ 60% RDX	50% TNT/ 50% TATB	50% TNT/ 50% NIGU
Atomic formula:			
C/100 g	2.01084	2.65961	2.00010
H/100 g	2.51062	2.27686	3.01228
N/100 g	2.17483	1.85074	2.58599
O/100 g	2.67852	2.48993	2.29503
Density, g cm ⁻³	1.69	1.73	1.65
Dimensions:			
diameter, mm	40	50	50
length, mm	90	90	90
Initiator type	cap 8	cap 8	cap 8
Booster type	10 g RDX	10 g RDX	10 g RDX
Charge weight, g	204.5	301.3	305.1
Dry soot weight, g	?	57	36
BKW calculations:			
CJ Pressure, kbar	261	239	230
CJ Density, g cm ⁻³	2.30	2.33	2.22
CJ Temperature, K	2,994	2,606	2,336

pressure (~300 kbar) regime led us to further characterize carbon in terms including such simple properties as morphology, phase (diamond or graphite), elemental composition and heat of formation. We examined the soot from the products of high explosives, well aware of the limitations of such recovery experiments; the products are recovered at ambient conditions, not at detonation pressure and temperature. But if diamond is seen, the diamond phase can be confidently associated with the higher pressures, although it can be questioned whether graphitic soot is formed in the high-pressure detonation state rather than in the release process.

Our examination of detonation soot is certainly not the first: see the work of Ornellas⁵ and Nomura *et al.*¹⁰ on pure TNT with initial density below the break. Using transmission electron microscopy (TEM), they saw graphitic structures similar to those reported here, although there was no indication of diamond clusters.

We describe soot recovered from the detonation of three explosive compositions fired in a 1,500-litre tank filled with 1 atm of argon. The explosive compositions were cast composites made from liquid TNT mixed with one of the following solids: cyclotrimethylenetrinitramine (RDX), triaminotrinitrobenzene (TATB), or nitroguanidine (NIGU) (see Table 1). After detonation, samples of the product gases were drawn for analysis, and the soot was recovered after the tank was emptied of the argon and detonation gases and brought up to atmosphere with air. The samples were oven-dried at 105 °C in air and stored in polyethylene bottles.

A small amount of each soot sample was prepared for observation by TEM. These samples were ultrasonically dispersed in acetone, and a commercial holey carbon film was dipped into the dispersion and air-dried. Dispersion quality was poor, so the preparation process probably altered the native agglomerate geometry significantly. Figure 1 shows a representative portion of the microstructure of sample 27 (60% TNT/40% RDX, approximately 'composition B'). Two distinct powder forms are present, one consisting of compact spheroids of ~7 nm diameter, the other of curved ribbons ~4 nm in thickness. In the image, the ribbons sometimes show the 0.35-nm interplanar spacing characteristic of both fully crystalline and turbostratically disordered graphites. The compact spheroids show no internal structure to the point resolution of our microscope (~0.35 nm). Because the two powder morphologies are agglomerated independently on a multi-micron scale, it is possible to explore their structures separately using convergent-beam electron diffraction. Figure 2 shows such a pattern obtained from a collection of the compact particles using a condensed beam of 20 nm diameter. The camera constant of the microscope was

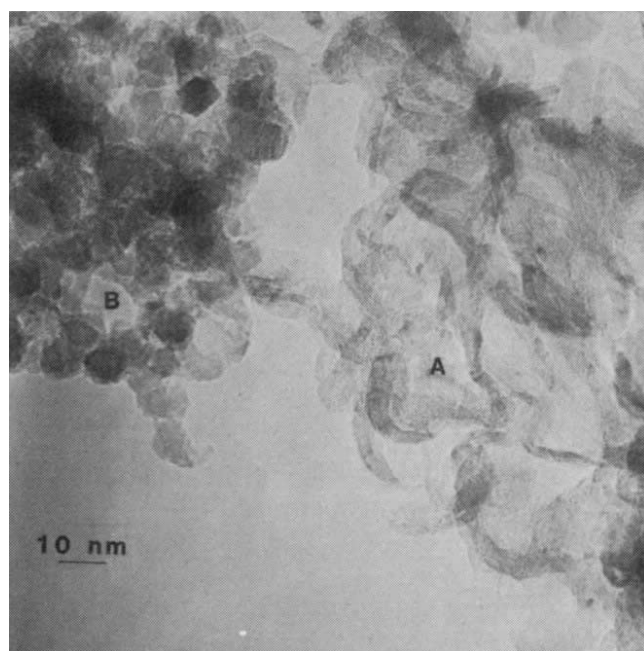


Fig. 1 Electron micrograph of sample 27, scale 10 nm. A, Graphite ribbons; B, diamond spheroids.

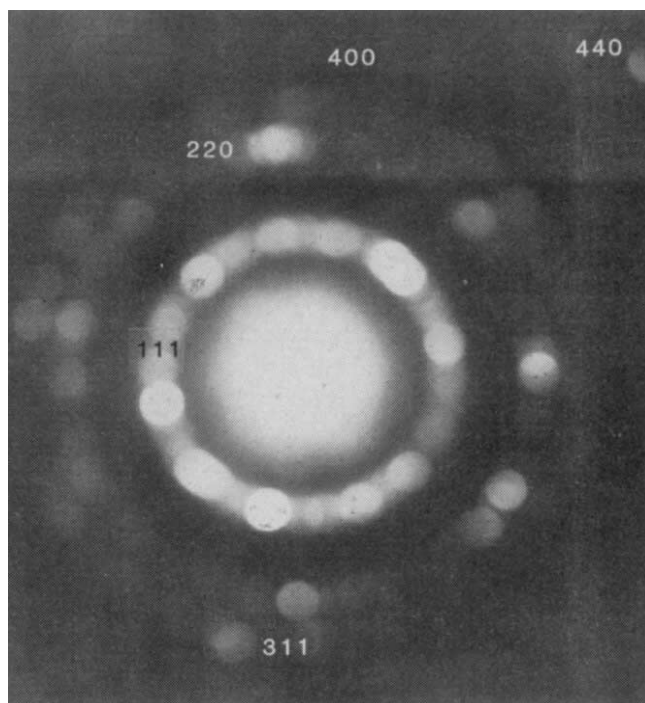


Fig. 2 Electron diffraction pattern showing diamond diffraction rings from (111), (220), (311), (440) and (400) weakly.

standardized against gold and SiC powders before and after the soot sample. The 'spotty ring' geometry of the pattern is expected from a crystalline sample whose grain size is not far smaller than the sampled area; the rings correspond to interplanar spacings of 0.2058, 0.1266, 0.1075, 0.884 and 0.636 nm respectively, in good agreement with the established (111), (220), (311), (400) and (440) spacings in diamond. These spacings do not agree with those expected in graphite.

Collimated-beam electron diffraction patterns from the turbostratic phase show two distinct intense rings, the inner corre-

sponding to the expected 0.35 nm (0001) spacing. Similar graphitic ribbons have been observed in soot from pure TNT¹⁰. The two other samples showed spheroids of 4 nm diameter and ribbons 2 nm thick.

For determination of the relative amounts of each of these morphologies in the soot, weighed amounts of sample 63 (50% TNT/50% NIGU) and sample 60 (50% TNT/50% TATB), whose TEM images showed a preponderance of spheroids, were fumed in nitric and perchloric acids to remove the graphite². In both cases, a brown powder was recovered (25% by weight), and X-ray diffraction powder patterns showed the same three diamond spacings mentioned above (0.206, 0.126 and 0.107-4 nm, matching the pattern from an authentic diamond sample) and a bandwidth consistent with crystals of ~4 nm in diameter. The patterns also showed the diffuse (002) band of graphite, which is attributed to amorphous carbon. The infrared spectrum of the residue from sample 63, with peaks at 2.94, 5.63, 6.13, 7.24 and 9.01 μm , closely matched the spectrum reported for 5-nm meteoritic diamonds, perhaps as a result of both samples being isolated by HClO_4 oxidation².

Now we turn to the interpretation of our observations. A recent theoretical presentation⁸ for carbon clustering in the detonation regime uses a model that is essentially as originally formulated for colloidal suspensions, with the rates for clustering dominated by diffusion. Because most of the energy release happens before cluster sizes reach 10^4 atoms, we are interested in larger carbon clusters, such as those in our soot, only to determine the nature of the complete energy release. But structural indications in the soot as to whether these clusters are fractal or compact, diamond or graphitic, guide us in our modelling. The TEM pictures indicate that we are viewing small carbon clusters, say less than 10^4 atoms, as compact 'blobs', spheres being a reasonable approximation for the diamond clusters and probably not a bad approximation for the graphitic clusters.

After two clusters collide and stick, we presume that an annealing process 'sphericalizes' the resulting cluster using the heat produced by the hot background fluid of N_2 , CO_2 , and H_2O and/or the energy released by the formation of bonds between the initial clusters. The annealing process shuts down when the fraction of atoms in the cluster surface is 10–20%: this is the 5-nm-sized cluster. Thus, we have a crude explanation for the basic cluster size of 5 nm.

Which combination of detailed mechanisms causes the termination of annealing is not known, but all have the same dependence on fraction of surface atoms. Some suggested mechanisms are: (1) the heat from bond formation becomes too small a fraction of the total energy of a cluster; (2) the effective melting temperature for a finite cluster shifts with cluster size; (3) the kernel in the coagulation equations⁸ changes character as a function of cluster size; and (4) the thermal fluctuations in a finite cluster are smaller on a percentage basis in larger clusters than in smaller clusters.

The morphology of the carbon is of interest because, to date, the modelling of carbon in the detonation environment has been with bulk equations of state, either graphite or diamond. Our data strongly indicate that we do not have bulk carbon. This soot is made up of clusters of either diamond or graphitic form, and the cluster sizes are small enough to have significant surface contributions. At earlier times during cluster growth⁸, the surface contributions would be even larger. Our diamond form differs from diamond produced by shocking graphite, which is dirty but bulk polycrystalline diamond.

There are interesting connections between our work and other recent experiments. The 5-nm diamond clusters recovered from meteorites are probably formed in a relatively high-hydrogen environment². Also, spherules in flame soots¹¹, are seen to be of similar size to those in the detonation soot and the astrophysical diamond clusters, although these are not diamond clusters in flames. Finally, in diamond thin-film technology, the desira-

bility of a high concentration of hydrogen in the vapour¹² is consistent with the hydrogen environment of the astrophysical diamonds and, as we shall argue, with the nitrogen environment of our detonation diamonds.

How can diamond form in the astrophysical and thin-film cases where the pressures are low, and in detonations where the surface energies will shift the diamond stability region to too high a pressure (see also refs 12 and 13). Carbon does not like unsatisfied bonds, and geometry could be used to close off the bonds, as for Buckminsterfullerene¹⁴. But there is another possibility in the various environments of interest here, for the astrophysical and thin-film conditions, there is hydrogen to cap off any dangling bonds, thereby shifting the diamond cluster stability back to lower pressures. For detonations, the hydrogen is tied up in water, but with nitrogen present, surface carbon bonds could be satisfied by coupling to CN functional groups. (In ref. 13, CO was discussed in this context, but we believe that CO is unfavourable compared with CN because oxygen bonding is more stable in H_2O and CO_2 .)

Lastly, we shift from the diamond clusters to the turbostratic graphite. Turbostratic graphite has the graphite structure in any given layer, but layers are random with respect to each other. As a result, the interlayer spacing increases from 0.335 nm to 0.344 nm (ref. 15), which is in good agreement with our observed 0.35 nm and reinforces our point that soot is not bulk graphite or diamond. Note that the turbostratic graphite is curved, and in that sense resembles Buckminsterfullerene.

The samples were provided by the Fraunhofer-Institut für Treib- und Explosivstoffe. We thank B. Roof for X-ray diffraction patterns, D. Hoard for the acid oxidations, and R. Lewis for an infrared spectrum of meteoritic diamond. We thank D. Schiferl, R. Ryan, D. Moore and M. S. Shaw for helpful discussions. The work is supported by the United States Department of Energy and the Department of the Army.

Received 25 January; accepted 27 April 1988.

1. Staver, A. M., Gubareva, N. V., Lyamkin, A. I. & Petrov, E. A. *Fizika Goreniya i Vzryva* **20**, 100–104 (1984).
2. Lewis, R. S., Ming, T., Wacker, J. F., Anders, E. & Steel, E. *Nature* **326**, 160–162 (1987).
3. Ree, F. H. *J. chem. Phys.* **84**, 5845–5856 (1986).
4. Ree, F. H. *J. chem. Phys.* **81**, 1251–1263 (1984).
5. Ornellas, D. L. *Colorimetric Determinations of Heat and Products of Detonation for Explosives: October 1961 to April 1982* (Lawrence Livermore Natn. Lab. report UCRL-52821, 1982).
6. Roy, R. *Nature* **325**, 17–18 (1987).
7. Mader, C. L. *Numerical Modelling of Detonations* 44–45 (University of California Press, Berkeley, 1979).
8. Shaw, M. S. & Johnson, J. D. *J. appl. Phys.* **62**, 2080–2085 (1987).
9. Urizar, M. J., James, E. Jr. & Smith, L. C. *Phys. Fluids* **4**, 262–274 (1961).
10. Nomura, Y. & Kawamura, K. *Carbon* **22**, 189–191 (1984).
11. Richter, R., Sander, L. M. & Cheng, Z. *J. Colloid Interface Sci.* **100**, 203–209 (1984).
12. DeVries, R. C. *A. Rev. Mater. Sci.* **17**, 161–187 (1987).
13. van Thiel, M. & Ree, F. H. *J. appl. Phys.* **62**, 1761–1767 (1987).
14. Heath, J. R., O'Brien, S. C., Curl, R. F., Kroto, H. W. & Smalley, R. E. *Accounts of Chemical Research* (submitted).
15. Kelly, B. T. *Physics of Graphite* 15–17. (Applied Science, London and New Jersey, 1981).

Solubility of colloidal ferric hydroxide

Lewis E. Fox

Department of Earth and Planetary Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

Considerable evidence supports the concept that phosphate^{1–4}, trace-metal cations^{5–7}, and rare-earth elements⁸ are strongly influenced by chemical interactions with suspended solids in turbid rivers and estuaries. Among the most active of suspended particulates are metal-hydroxide colloids, which exist as dispersions or as surface coatings on larger suspended solids. Elucidation of the chemical mechanisms governing colloidal and dissolved species interactions is hampered by ignorance concerning fundamental chemical parameters describing the colloidal phases. Colloidal

ferric hydroxides have long been known to interact with a variety of dissolved species in natural waters, yet reported solubility products for ferric hydroxides range over three orders of magnitude⁹. Without such fundamental information, more refined issues, such as the extent to which solid solution formation governs dissolved species, are impossible to resolve. Here we re-examine the solubility properties for colloidal ferric hydroxide reported over the past 60 years, and combine them with results from experiments with dialysed ferric-hydroxide suspension performed in this laboratory. After correction of historical data for ferric complexation and variations in ionic strengths, the combined data are consistent with the mass action equation,

$$\alpha_{\text{Fe}^{3+}} \alpha_{\text{OH}^-}^{2.35} = 10^{-31.7} \quad (1)$$

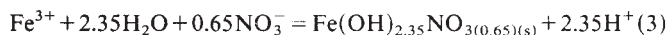
where α_i is the activity of species i .

The activities of ferric ion calculated from measurements made in this laboratory, as well as those reported in the literature (see Table 1), are plotted as a function of pH in Fig. 1. Clearly there is consistency between measurements made from dialysed solutions and those obtained from the literature. The scatter of points from previous work, which fall significantly below the line, mostly represent data from hydrolysed ferric sulphate solutions, which form predominant FeSO_4^+ complex instead of colloidal ferric hydroxide¹⁰. The least-squares fit to the data summarized in Fig. 1 yields a correlation coefficient of 0.997 and fits the equation,

$$-\log \alpha_{\text{Fe}^{3+}} = 2.35 \text{ pH} - 1.17 \quad (2)$$

where α_i represents the activity of species i .

This result implies that the solid phase has a constant composition over the entire range of pH values studied and the reaction leading to the formation of ferric hydroxide can be expressed as



Using $K_{\text{water}} = 10^{-14}$, the solubility of ferric hydroxide can thus be expressed through the mass action law as equation 1. By convention, species belonging to the solvent, H_2O and NO_3^- , are omitted from the expression because their concentrations are constant^{10,11}. Nitrate ion provides charge balance but does not seem to effect the solubility of colloidal ferric hydroxide, perhaps because ferric-hydroxide polycations do not include the counterion in the polymer chain¹⁰.

The means and standard deviations of the solubility product and the OH:Fe compositional ratio determined from the dialysis data (literature data excluded) are 31.7 ± 0.37 and 2.33 ± 0.07 , respectively. The values were calculated from a data set consisting of 38 separate measurements. At the 99% confidence interval, the range of solubility product values and compositional ratios are 31.7 ± 0.13 and 2.33 ± 0.03 , respectively. The compositional ratio (2.35) determined from the least-squares fit of the data shown in Fig. 1, which includes measurements from the literature, is not significantly different from the values determined exclusively from the dialysed solutions.

Reported ion products measured from suspensions of colloidal ferric hydroxide prepared by hydrolysis of ferric salts are summarized in Table 1, along with corresponding ion activity products corrected for both the presence of ferric-hydroxide complexes and for ionic strength. The measurements obtained from the literature were made by using either potentiometric titration¹²⁻¹⁶ or redox electrode^{11,17,18}. Ion activity values derived from OH:Fe ratios of both 3 (column 3) and 2.35 (column 4) are listed for comparison. It is evident from Table 1 that the historical data, once corrected for ionic strength and complex formation, show relatively constant ion activity products, but the techniques used to obtain the cited values are applicable only at pH values between 2 and 3, and at elevated concentrations of suspended iron. Once the range of pH values is extended,

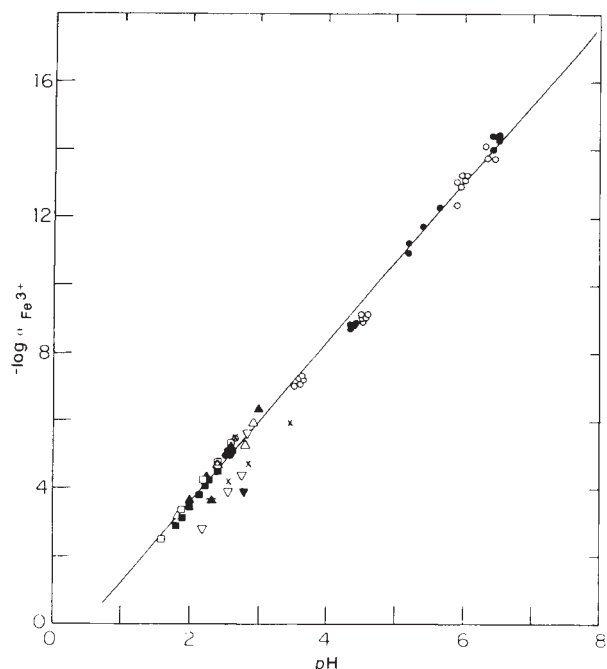


Fig. 1 The negative logarithm of ferric ion activity as a function of pH for colloidal ferric hydroxide prepared from various ferric salts. The correlation coefficient is 0.997 and the slope is 2.35 ($n=60$, excluding seven outliers). The solubility product taken from this graph is $10^{-31.7}$. Cited ion concentrations were corrected for ionic strength by the mean salt method²³. Those published before 1949 were also corrected for the presence of ferric hydroxide complexes. The following symbols represent activities calculated from measurements of ferric hydroxide reported in the literature: Δ , ref. 12; $\star$, ref. 13; $\times$, ref. 14; ∇ , ref. 16; $\blacktriangledown$, ref. 15; $\blacktriangle$, ref. 17; $\blacksquare$, ref. 18; $\triangle$, ref. 11; $\otimes$, ref. 9. Open circles ($\circ$) represent samples taken from within a 1,000-dalton dialysis membrane with an external ferric solid concentration of 100 μM . Filled circles ($\bullet$) represent samples taken from within a 1,000-dalton dialysis membrane with an external ferric solid concentration of 1,000 μM .

Methods. Dialysed solutions were prepared from a suspension of 3 mM colloidal ferric hydroxide which was produced by dissolving ferric nitrate in de-ionized water and allowing the solution to react for three months. Ferric hydroxide suspensions prepared in this manner are reported to contain 30–35% colloidal goethite and 65–70% ferric polycations¹⁰, which are stable for at least 2 yr (ref. 24). Concentrations of hydrogen ion in the ferric-hydroxide suspensions were adjusted with sodium hydroxide to obtain a range of pH values between 2.5 and 6.5. Pre-washed dialysis tubing that excludes solids with molecular weights greater than 1,000 daltons was filled with de-ionized water and immersed in the ferric-hydroxide solutions. Experimental solutions were continuously stirred with magnetic stir-bars. Samples were taken daily from within each dialysis tube for two weeks. Concentrations of total iron were determined by atomic absorption spectroscopy using standard methods and a graphite furnace. Hydrogen ion activity was measured potentiometrically with a solid-junction, glass electrode fitted to a Beckman, Omega 71, pH meter. Activity calculations are given elsewhere²⁵. The possibility was considered that monomeric, dimeric and trimeric ferric-hydroxide, ferric-carbonate and ferro-organic complexes could cross the dialysis membrane. Published equilibrium constants for ferric-hydroxide complexes indicate that only monomeric species exist in significant concentrations in the range of pH values considered here, although the colloidal behaviour of ferric hydroxide suggests the existence of larger, as yet unquantified complexes²⁶. Carbon analysis of the stock ferric nitrate indicated <0.01% organic contamination. The low organic concentrations prevent interference from ferro-organic complexes. The possible influence of ferric carbonate complexes is unknown, as stability constants for these substances are not readily available.

Table 1 Solubility parameters for colloidal ferric hydroxide

Reference	Reported ion product* $p([\text{Fe}^{3+}] \cdot 3[\text{OH}^-])$	Reported ion activity product $p(\alpha_{\text{Fe}^{3+}} \cdot 3\alpha_{\text{OH}^-})$	Ion activity product† $p(\alpha_{\text{Fe}^{3+}} \cdot 3\alpha_{\text{OH}^-})$	Ion activity product† $p(\alpha_{\text{Fe}^{3+}} \cdot 2.35\alpha_{\text{OH}^-})$	Temp. (°C)	Counterion concentration	Age	pH (Range of measurements)
This paper	—	—	36.6–39.2	31.7 ± 0.3	25	0.05 M NO_3^-	3 months	2.6–6.6
12	37.9	—	39.4 ± 0.5	31.9 ± 0.2	14	0.5 M Cl^-	F‡	2.7–3.3
13	37.7	—	39.0 ± 0.1	31.0 ± 0.1	18	0.15 M Cl^-	F	2–3
14	—	36.5 ± 1.1§	38.3 ± 0.6	31.0 ± 0.3	—	0.05 M Cl^-	F	2.7–3.6
16	35.6 ± 1.4	—	38.7	31.2	18	0.05 M SO_4^{2-}	F	2.3–3.0
22	—	39.2	—	—	—	—	—	—
15	35.5	—	37.5	30.2	20	0.1 M $\text{FeNH}_4(\text{SO}_4)_2$	F	2–3
17	38.7	—	39.7 ± 0.2	31.9 ± 0.2	25	3 M NaClO_4	8 days	1.7–2.7
18	39.1	—	39.4 ± 0.2	31.7 ± 0.1	25	3 M NaClO_4	8 days	1.7–2.9
11	38.8	—	39.6 ± 0.3	32.0 ± 0.2	25	3 M NaClO_4	3 weeks	1.9–3.0
9	38.5	—	39.5	32.1	25	0.1 M SO_4^{2-}	3 days	2.65
Means			39.0 ± 0.7	31.4 ± 0.6				

* p denotes the negative logarithm.

† Activity coefficients for Fe^{3+} were determined by the mean salt method²³. Measurements before 1959 were also corrected for the presence of $\text{Fe}(\text{OH})^{2+}$ and $\text{Fe}(\text{OH})^+$.

‡ F denotes freshly precipitated colloidal ferric hydroxide.

§ Activity coefficients were estimated by the original authors from expanded Debye–Hückel theory to be 0.2. Mean salt estimation of the coefficient is 0.196.

the OH:Fe ratio of 3 produces a wide range of ion activity products (Table 1, column 3), suggesting the ratio of 2.35 to be more acceptable.

The mean of OH:Fe ratios reported in ref. 19 is 2.41 ± 0.13 for ferric polycations isolated from hydrolysates of ferric nitrate. This observation supports our results and tends to suggest that ferric polycations govern the solubility of ferric nitrate hydrolysates, at least in the short term, as the presence of colloidal goethite, which could be expected to comprise 30–35% of our suspensions¹⁰, does not seem to significantly influence the results. Ferric polycations are characterized by linear polymers joined by double hydroxide bridges. If polycation solubility dominates, then the solubility of hydrolysates of ferric chloride and ferric perchlorate would be expected to be independent of the varying amounts of colloidal $\beta\text{-FeOOH}$ and goethite that form with ferric polycations during hydrolysis of those salts^{10,20–22}. This view is consistent with the adherence of most values taken from literature data, which were derived primarily from hydrolysis of ferric chloride and perchlorate, to the slope generated by hydrolysates of ferric nitrate (Fig. 1). Thus, with the exception of ferric sulphate, the solubility parameters reported here may be reasonable approximations for most colloidal ferric hydroxide suspensions.

I thank Fanya Gimmelfarb and Amy Snodgrass for their assistance in the laboratory, and Dr Steven Wofsy for helpful suggestions. This work was supported by NSF and EPA grants to Harvard University.

18. Schindler, von P., Michaelis, W. & Feitknecht, W. *Helv. chim. Acta* **46**, 444–449 (1963).
19. Murphy, P. J., Posner, A. M. & Quirk, J. P. *J. Colloid. Interface Sci.* **56**, 284–297 (1976).
20. Musić, S., Nagy-Czako, I. & Vertes, A. *Colloid. Polymer Sci.* **258**, 469–470 (1980).
21. Crosby, S. A., Glasson, D. R., Cutler, A. H., Butler, I., Turner, D. R., Whitfield, M. & Millward, G. E. *Envir. Sci. Technol.* **17**, 709–713 (1983).
22. Cooper, L. H. N. *Proc. R. Soc. London* **B124**, 299–307 (1938).
23. Millero, F. J. & Schreiber, D. R. *Am. J. Sci.* **282**, 1508–1540 (1982).
24. Murphy, P. J., Posner, A. M. & Quirk, J. P. *J. Colloid. Interface Sci.* **56**, 270–283 (1976).
25. Fox, L. E. *Geochim. cosmochim. Acta* (in the press).
26. Matijević, E. & Janauer, G. E. *J. Colloid. Interface Sci.* **21**, 197 (1966).

Throughflow into the Indian Ocean through the Lombok Strait, January 1985–January 1986

Stephen P. Murray & Dharma Arief

Coastal Studies Institute and Department of Geology and Geophysics, Louisiana State University, Baton Rouge, Louisiana 70803, USA

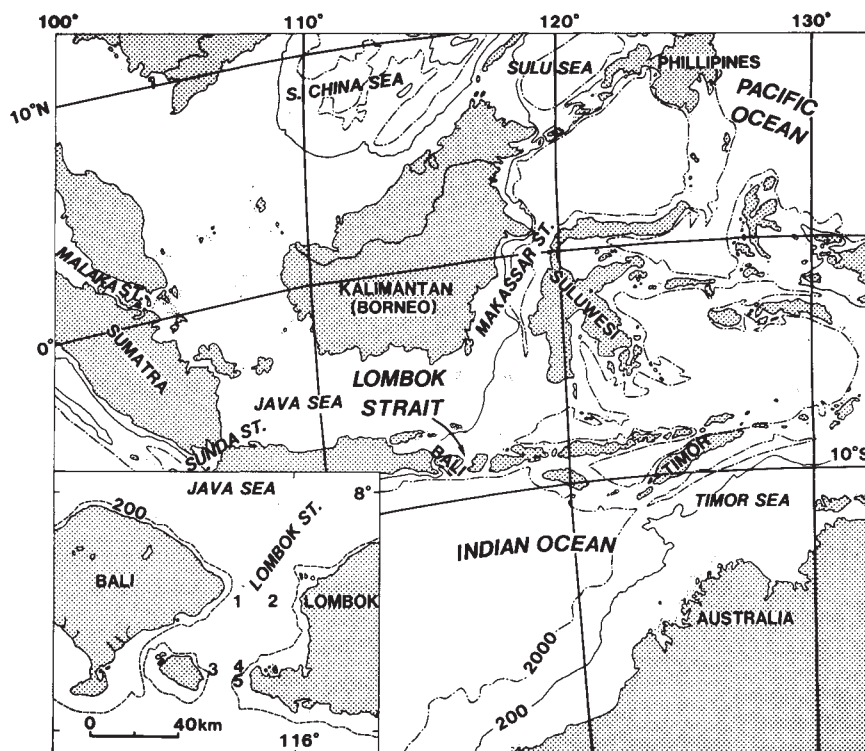
Recent papers (for example, refs 1 and 2) have raised the possibility, perhaps of global significance, that the rate of interbasin mass exchange between the Pacific and Indian Oceans through the Indonesian archipelago (referred to as the throughflow) is far higher than the value of 1–2 Sv (1 Sv = $10^6 \text{ m}^3 \text{ s}^{-1}$) originally suggested in ref. 3. The elevated temperature and depressed salinity of throughflow water have a critical role in the heat and freshwater balance of the Indian Ocean⁴. Inasmuch as the Indonesian throughflow route is the only interocean basin connection in tropical latitudes, its presence and strength have important implications to heat flux through the global ocean⁵. Present estimates of the magnitude of the throughflow are all by various indirect methods and vary greatly, from 1.5 to 20 Sv (refs 1–3, 6–8; reviewed in ref. 5). We present here direct (current meter) observations that support the existence of a large throughflow.

A distinct annual cycle in the observed transport shows a broad minimum of ~1 Sv from February to May inclusive during the west monsoon and spring transition and a maximum of ~4.0 Sv in the northern summer months of the east monsoon. Our observed mean annual transport of 1.7 Sv for the Lombok

Received 11 December 1987; accepted 1 March 1988.

1. Chase, E. M. & Sayles, F. L. *Estuar. and Coast. mar. Sci.* **11**, 383–391 (1980).
2. Smith, D. J. & Longmore, A. R. *Nature* **287**, 532–534 (1980).
3. Fox, L. E., Sager, S. L. & Wofsy, S. C. *Limnol. Oceanogr.* **30**, 826–832 (1985).
4. Fox, L. E., Sager, S. L. & Wofsy, S. C. *Geochim. cosmochim. Acta* **50**, 783–794 (1986).
5. Farley, K. J., Dzombak, D. A. & Morel, F. M. M. *J. Colloid Interface Sci.* **106**, 226–242 (1985).
6. Davis, J. A., Fuller, C. C. & Cook, A. D. *Geochim. cosmochim. Acta* **51**, 1477–1490 (1987).
7. Comans, R. N. J. & Middleburg, J. J. *Geochim. cosmochim. Acta* **51**, 2587–2591 (1987).
8. Goldstein, S. J. & Jacobsen, S. B. *Isotope Geosci.* (in the press).
9. Langmuir, D. & Whittemore, D. O. *Adv. Chem. Ser.* **106**, 209–234 (1971).
10. Musić, S., Vertes, A., Simmons, G. W., Nagy-Czako, I. & Ledihseier, H. *J. Colloid. Interface Sci.* **85**, 256–266 (1982).
11. Biedermann, G. & Chow, J. T. *Acta chem. scand.* **20**, 1376–1388 (1966).
12. Jellinek, K. & Gordon, H. Z. *Physik. Chem.* **12**, 207–249 (1924).
13. Britton, H. T. S. *J. chem. Soc.* **127**, 2148–2159 (1925).
14. Elder, L. W. *Trans. electrochem. Soc.* **57**, 383–393 (1930).
15. Evans, U. R. & Pryor, M. J. *J. Chem. Soc. (London)* **5**, 157–160 (1949).
16. Kriukov, P. A. & Awsejwitsch, G. P. Z. *Elektrochem.* **39**, 884–891 (1933).
17. Biedermann, G. & Schindler, von P. *Acta chem. scand.* **11**, 731–740 (1957).

Fig. 1 Map of the Indian-Pacific convergence region with an inset of the Lombok Strait showing sites of five current-meter moorings, located with numbers. Water depths at each site are: (1) 1,060 m; (2) 1,080 m; (3) 450 m; (4) 450 m; (5) 150 m.



Strait alone supports the higher values of throughflow (10–15 Sv) recently deduced^{1,2} for the entire archipelago. The annual variation of the Pacific-Indian throughflow predicted by the NORDA Global Ocean Model (J. Kindle *et al.* in preparation) is in phase with our observed transport and suggests that the Lombok Strait may carry 20% or more of the total throughflow.

Despite numerous breaches in the archipelago, the deep passages (~2,000 m depth) bracketing Timor were considered³ as the sole pathways for any significant transport into the Indian Ocean. In terms of cross-sectional area available for transport, the second most important channel through the lower archipelago is the Lombok Strait between the islands of Bali and Lombok (Fig. 1). Lying at the end of a deep bathymetric trough linking it to the Makassar Strait, it provides a direct conduit for Pacific Ocean water into the Indian Ocean.

The large-scale meteorological forcing is dominated by the southeast Asian monsoon, which brings strong east winds across the archipelago from May to early September and west winds from November to March. Transition periods are characterized by weak, variable winds. A persistent southerly surface flow down the Makassar Strait throughout the year^{3,9} suggests a regional pressure gradient directed toward the Indian Ocean upon which any monsoonally induced oscillations would be superimposed.

The Lombok Strait exhibits depths of 800–1,000 m except at the south end, where a small island (Nusa Penida) splits the entrance. Sills cross the channels on each side of the island, with maximum depths of ~350 m. The data from five arrays of current meters (subsurface taut-line moorings located in the inset to Fig. 1) are used to determine the net transport through the Strait in 1985 and its seasonal variability. Our initial data from mooring 5, on the flank of the sill in the south strait, showed tidal current speeds of $\pm 3.5 \text{ m s}^{-1}$. Thus the drag forces at this minimum-area cross-section were too severe for a picket line of taut-line moorings. In the north strait, however, tidal current amplitudes range from 20 to 50 cm s^{-1} over the spring-neap cycle. Accordingly we maintained north strait moorings 1 and 2 through the year and instrumented sites 3 and 4–5 alternately on the first two of our three deployments to study spatial gradients and sill effects. EG&G acoustic releases and Endeco

174 (impeller) and 1174 (rotor) current meters, recording at $\Delta t = 8 \text{ min}$, proven successful in other straits¹⁰, yielded 93 months of good-quality current meter data.

An example of our observations of the vertical structure of the mean flow in the north strait, site 2, during the first deployment (Fig. 2) suggests a persistent mean flow to the south, with a strong decrease in speed between the 75-m and 300-m levels. The episodic flow reversals in the upper level are associated with intense cyclones crossing the Timor Sea, which will be discussed in detail elsewhere. The observed vertical extent of the mean flow agrees quite well with the only other available data, from the Snellius anchor station in the Makassar Strait¹¹, which show a strong ($50\text{--}60 \text{ cm s}^{-1}$) mean southward flow concentrated in the upper 200 m.

A 13-month time series of the low-pass currents at 35 m depth (Fig. 3) demonstrates the seasonal variability of the mean current through the year. Except for reversals during the cyclone season, southerly components dominate throughout the year. Maximum speeds are reached from July to September during the peak of the east monsoon, and a broad minimum occurs from October through the rest of the year.

The volume transport T through the Lombok Strait across $8^\circ 27' \text{ S}$ is computed for each month using the monthly means of the north-south component of the low-pass current from mooring sites 1 and 2 and the cross-sectional area, determined from our bathymetry profiles and USNO Chart 72222. The shape of the velocity profile between the 75-m level and the 300-m level is interpolated using the Snellius data¹¹ as a template and is consistent with our own 200-m-level observations at sites 3 and 4. The speed profiles are resampled at 50-m increments and the transport computed from

$$T = \sum_{j=1}^{j=2} \sum_{i=1}^{i=7} v_{ij} b_{ij} \Delta z$$

where $j = 1$ (or 2) indicates summation over the west (or east) half of the cross-section with velocity data from mooring site 1 (or 2). The i index performs vertical summation of the velocity component v and the channel half width b (both depth dependent) over a constant vertical increment $\Delta z = 50 \text{ m}$.

We find that ~80% of the net transport is contained in the

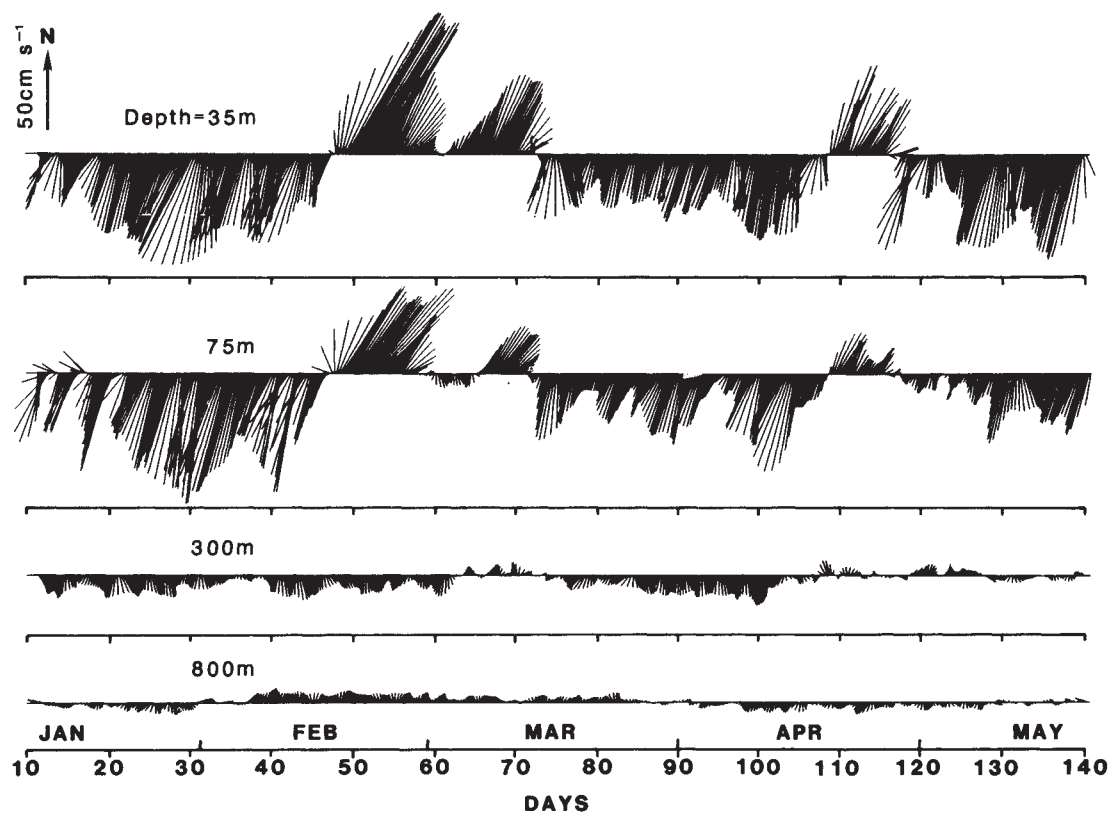


Fig. 2 Current velocity vectors (60-h low-pass) from the four levels of the north strait mooring (site 2) from January to May 1985. The area at the north-strait cross section above the 400 m depth level is $13.3 \times 10^6 \text{ m}^2$.

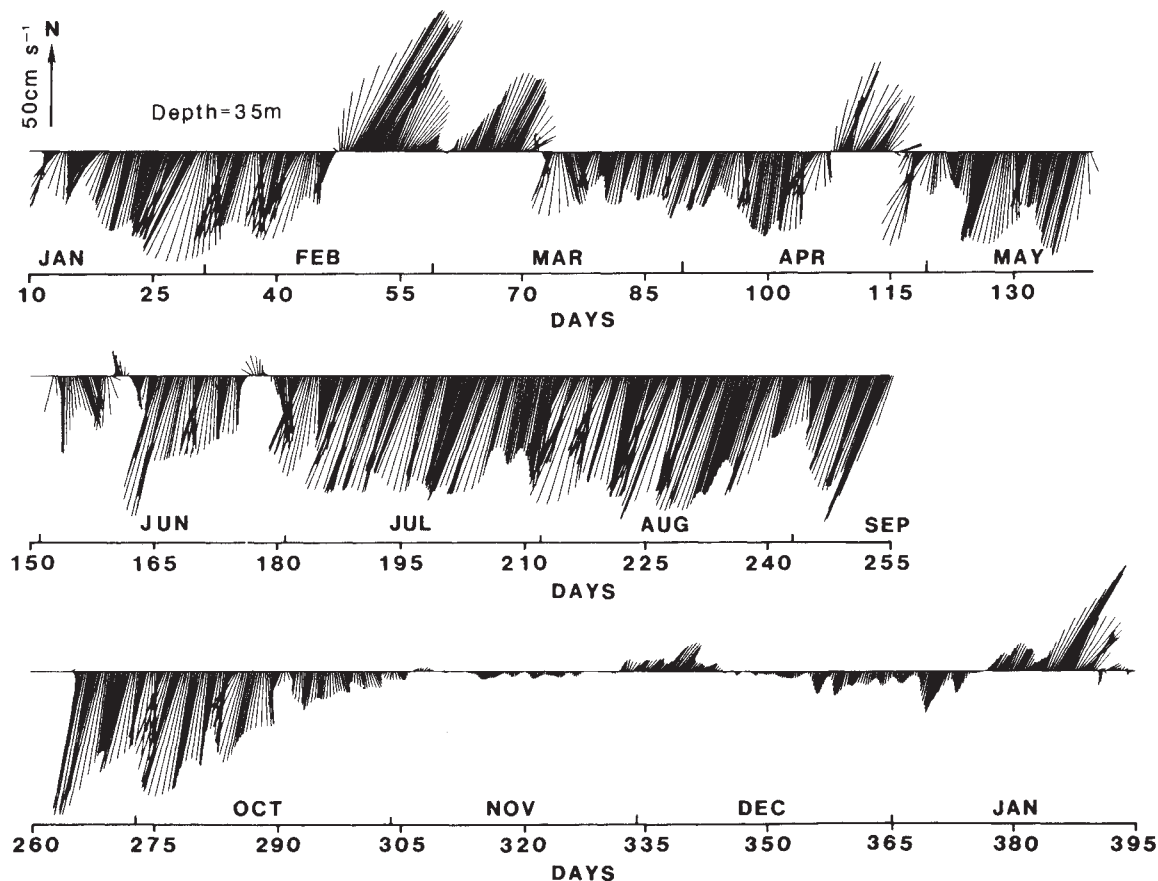


Fig. 3 Thirteen-month time series of current vectors from the 35 m depth level at the site 2 mooring.

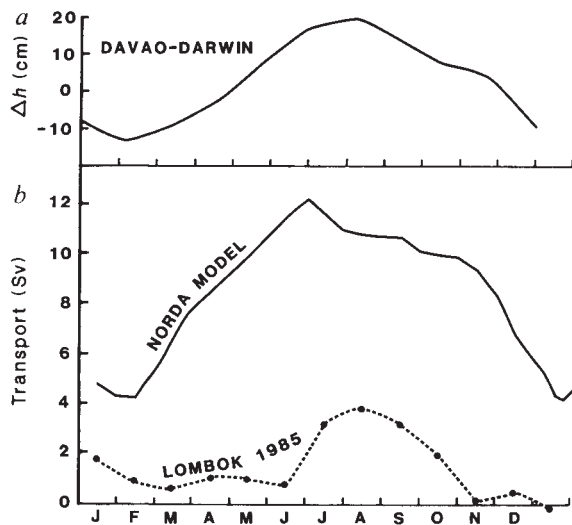


Fig. 4 a, Average annual cycle for 1978-83 of the sea level drop from Davao, Philippines, to Darwin replotted from ref. 12. b, Dotted line, annual cycle of monthly averaged transport through the Lombok Strait computed from measurements during January 1985 to January 1986. Solid line, annual cycle of the Pacific to Indian throughflow across 120° E from 10° S to 18° 45' S (from J. Kindle *et al.*, in preparation).

upper 200 m, in good agreement with Wyrtki's¹² conclusion that the pressure gradient driving the throughflow is similarly concentrated in the upper 200 m. The internal Rossby radius in the north strait is 90 km, three times the width of the Strait, indicating a dynamically narrow channel¹³. Furthermore, we neglect the transport in lateral frictional wall layers of thickness $(\nu_H/f)^{1/2}$ (ref. 14), here ~ 1 km, where ν_H (the lateral eddy viscosity) = $0.0103 l^{1.1} = 1.4 \times 10^5 \text{ cm}^2 \text{ s}^{-1}$ (ref. 15); the length scale l is taken as the strait width (30 km) and f is the Coriolis parameter.

The results of the transport computations (Fig. 4) show a distinct annual cycle with minimum transport in early 1985 of ~ 1 Sv during February-May (west monsoon and spring transition), a maximum of 3.8-4 Sv in July and August (east monsoon), followed by a deeper minimum at the end of the year. The 1985 average transport is 1.7 Sv, with a seasonal r.m.s. deviation of 1.2 Sv. These data suggest that the transport into the Indian Ocean through the Lombok Strait alone is as large as was previously attributed to the entire archipelago³.

None of our extensive data indicate that the net southward transport in the north strait is being blocked by the sill to form a deep northward return flow. For example, (1) in the north strait when net speeds at the 35-m level are high to the south, speeds at 500 m and 800 m depths are also to the south at 0.02-0.04 m s^{-1} for long periods; (2) 6 km north of the sill at mooring sites 3 and 4, net speeds at the 200-m level are also southward at 0.15-0.20 m s^{-1} during the largest transport in July and August; (3) net currents on the sill at site 5 from January to May 1985 are consistent with the total net transport observed simultaneously in the north strait. In addition, our data from salinity-temperature-depth (STD) casts clearly indicate waters above the 300-m level flow south over the sill into the Indian Ocean. There is no indication of a downwelling northward of the sill. On the contrary, the distinct core layer of the Northern Subtropical Lower Water between the 125 m and 175 m depths is clearly tracked in January, June and September 1985 (ref. 16) through the Strait and over the sill into the Indian Ocean.

As shown in ref. 17, a nonlinear tidal residual circulation in the north strait that might bias our throughflow measurements should be modulated by the spring-neap tidal cycle, which is strong in the strait. Isolating the current in the fortnightly tidal band (13-15 days) at the upper level current meters where

transport is strongest yields an irregular signal with a maximum amplitude of about 6% of the total low pass current, indicating that this effect is negligible.

Also plotted in Fig. 4 is the Pacific to Indian Ocean transport predicted by the reduced-gravity Global Ocean Model of the Naval Ocean Research and Development Activity (NORDA) (J. Kindle *et al.*, in preparation). For earlier model results, see ref. 18. The annual cycle observed in the Lombok data, with maximum throughflow during June-August, at the peak of the east monsoon, is clearly present in the NORDA model data, showing minimum and maximum throughflows into the Indian Ocean of 4.5 and 12 Sv respectively. The annual mean is 8.9 Sv, with a seasonal amplitude of 3.8 Sv, suggesting that the Lombok Strait is carrying at least 20% of the interbasin transport. The initial model did not include the Lombok Strait, but this study has provoked its incorporation, and direct comparisons of transport will soon be possible.

It is clear from the model results that the annual cycle in the Lombok Strait component of the throughflow is related to the annual cycle of the monsoon winds. The annual cycle in the sea level difference between the Philippines and the Indian Ocean at Darwin, presented in ref. 12, is replotted in Fig. 4. It shows a maximum sea-level drop in the northern summer and a minimum in the northern winter, in phase with our Lombok transport observations, supporting Wyrtki's conclusion^{3,12} that a sea-level slope from the Pacific to the Indian Ocean is involved in driving a net transport into the Indian Ocean, although the magnitude of this transport remains in question.

We thank R. Fredericks, W. Winans and F. DeMers for superb technical support under less than ideal conditions, and J. Kindle for use of the most recent NORDA model data before publication. This work was supported by Coastal Sciences Program of the Office of Naval Research, Arlington, Virginia.

Received 14 December 1987; accepted 25 March 1988.

- Godfrey, J. & Golding, T. *J. phys. Oceanogr.* **11**, 771-779 (1981).
- Pirola, A. & Gordon, A. *J. phys. Oceanogr.* **14**, 747-753 (1984).
- Wyrtki, K. *NAGA Report 2* (1961).
- Toole, J. & Raymer, M. *Deep Sea Res.* **32**, 917-928 (1985).
- Gordon, A. *J. geophys. Res.* **91**, 5037-5046 (1986).
- Cox, M. *Numerical Models of Ocean Circulation* (National Academy Press, Washington, DC, 1982).
- Fine, R. *Nature* **315**, 478-480 (1985).
- Fu, L. *J. phys. Oceanogr.* **16**, 1683-1693 (1986).
- United States Naval Oceanographic Office, Spec. Pub. 1404-1N 5 (NSTL Station, Michigan, 1977).
- Murray, S., Hecht, A. & Babcock, A. *J. mar. Res.* **42**, 265-287 (1984).
- Lek, L. *Die Ergebnisse der Strom und Serienmessungen, The Snellius Expedition Vol. II, Part 3* (Brill, London, 1938).
- Wyrtki, K. *J. geophys. Res.* **92**, 12941-12946 (1987).
- Gill, A. *Atmosphere-Ocean Dynamics* (Int. Geophys. Ser., 30) (Academic, New York, 1982).
- Killworth, P. *J. phys. Oceanogr.* **3**, 3-15 (1973).
- Okubo, A. Tech. Rept. 38, Chesapeake Bay Inst. (Johns Hopkins Univ., Baltimore, 1968).
- Arief, D. & Murray, S. *Eos* **68**, 1746 (1987).
- Tee, K. T. *J. phys. Oceanogr.* **7**, 396-402 (1977).
- Kindle, J., Heburn, G. & Rhodes, R. in *Further Progress in Equatorial Oceanography* (eds Katz, E. J. & Witte, J. M.) 317-321 (Nova Univ. Press, Fort Lauderdale, 1987).

Importance of crystal settling in the differentiation of Deccan Trap basaltic magmas

K. G. Cox & C. Mitchell

Department of Earth Sciences, Oxford University, Parks Road, Oxford OX1 3PR, UK

Recent research on the mechanism of differentiation of magmas has seriously challenged the idea of crystal settling, and invited alternative interpretation in terms of convective fractionation. Here we present data for basalts that suggest that the older hypothesis should still be regarded as significant in the evolution of this

magma type. We describe mixed sequences of aphyric and porphyritic eruptions among which only the denser liquids have brought suspended crystals to the surface, discriminating strongly in favour of plagioclase, the least dense solid phase. By implication, ferromagnesian minerals have been selectively lost by crystal settling from the porphyritic magmas, and all crystallizing phases have been lost from the less dense aphyric magmas.

First discussed in detail by Darwin¹, the crystal-settling hypothesis, depending on density contrasts between liquids and crystals, became well established as an important but not exclusive mechanism of magmatic differentiation during the twentieth century with the study of layered intrusions (see ref. 2 for review). But serious doubts about its efficacy began to emerge in the 1970s from detailed studies of plutonic rocks^{3,4}, from the demonstration that plagioclase probably could not sink in Fe-rich basaltic liquids⁵, and from experiments suggesting that silicate melts were fluids with finite yield strengths⁴.

Additionally, experimental work on convection in magma chambers⁶ indicated that magmatic flow was commonly likely to be vigorous enough to suppress crystal settling. At times during the past decade some earth scientists have been sufficiently impressed by the objections to crystal settling to deny its role as an important mechanism. Sparks *et al.*⁶, for example, postulate a mechanism termed convective fractionation as the major factor controlling differentiation. In this, crystallizing material grows on the magma chamber margins, and residual liquid convects away from it.

We believe, however, that recent debate on mechanisms of differentiation has underestimated the importance of crystal settling, and has encouraged many to ignore older, and by no means ill-founded, work.

Here we discuss data from certain common types of extrusive rocks, an important distinction from studies that have concentrated on the intrusive record. Volcanic rocks, unlike plutonics, give us different snapshots in time during magmatic evolution. While the study of plutonic rocks enables direct examination of crystallization products formed *in situ* on magma-chamber margins, it gives only indirect evidence of the formation of crystals that we term *within-chamber* that is, growing within the liquid inside the chamber, and remaining there unless removed by sinking/floating mechanisms or by eventual adherence to the chamber walls. Phenocryst-bearing lavas, we assume, give direct information about within-chamber crystallization products.

For our study we have chosen the Ambenali Formation of the Deccan Traps of India^{7,8}, a formation consisting of ~500 m of tholeiitic basalt flows, which because of extreme isotopic uniformity are generally regarded as virtually free of crustal contamination. Within the formation, sub-sequences of varying petrographic type, grading from completely aphyric to mildly porphyritic (up to ~25% of phenocrysts of olivine, clinopyroxene and plagioclase), can be traced for distances of up to 100 km. We shall attempt to interpret the relationship between a set of aphyric samples and a set of relatively highly porphyritic samples (selected as having >10 vol.% of phenocrysts).

The two groups define only slightly overlapping compositional trends (Fig. 1), the porphyritic being displaced towards Mg-poor compositions. The Fe-enrichment with decreasing MgO is typical of tholeiitic fractionation, generated by the removal of a gabbroic mineral assemblage, and has been described before (for instance in ref. 9) from the Deccan. We first interpret the aphyric trend in terms of the fractionation of such an assemblage.

Mixing calculations¹⁰ have been used to relate a generalized primitive parent (Fe-poor) to a generalized Fe-rich daughter product by the extraction of the three phases. As it is difficult to judge the mineral compositions to use, we have made many calculations based on a range of mineral compositions. Figure 2 gives the results and shows that fractionation in the approximate proportions of 10–15% olivine, 30–40% clinopyroxene and 50–60% plagioclase is required. This is similar to previous

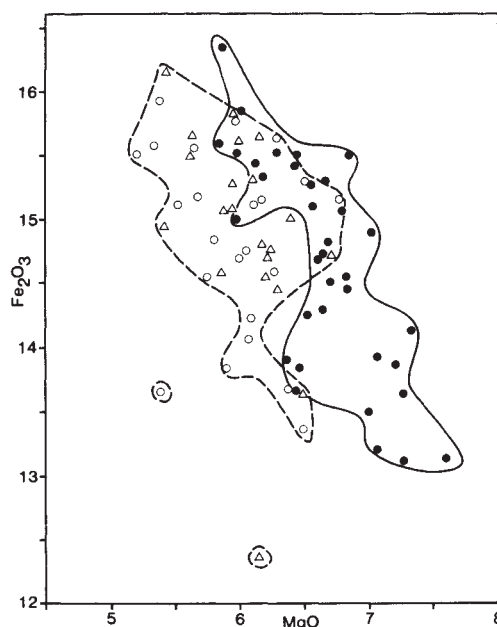


Fig. 1 MgO plotted against Fe_2O_3 (total) for Ambenali Formation basalts. Filled symbols, aphyric rocks; open circles, modally analysed porphyritic rocks with >10 vol. % phenocrysts; open triangles, rocks in which visually estimated modes include >10% phenocrysts. Each modal analysis and visual estimate is based on two standard thin sections. The estimates are thought to be accurate to ± 3 vol. %. Of the chemical analyses, 36 are from Devey¹¹, the remaining 42 are new. Devey¹¹ has also provided 9 of the 22 modal analyses.

estimates of cotectic proportions during Deccan basalt fractionation¹¹.

Figure 2 also shows the compositions of liquids saturated with olivine, clinopyroxene and plagioclase, within the simplified normative basalt tetrahedron¹² under low-pressure dry conditions (though this projection is not strongly pressure-sensitive within the range of crustal pressures). As the diagram has many of the properties of a ternary phase diagram, the position of the cotectic liquids also indicates the proportions in which the phases should crystallize. The near coincidence of the cotectic liquids with the calculated extracts indicates that the latter are consistent with the observed fractionation of bulk compositions and with the phase relations. The aphyric rocks themselves, assumed to represent the natural cotectic liquids, also plot close to the cotectic area, and close to the calculated extracts. We conclude that they represent magmas that originated by the fractionation of olivine, clinopyroxene and plagioclase, in cotectic proportions at a relatively low pressure.

Comparison with the phenocryst proportions in the porphyritic rocks, however, reveals a notable discrepancy. Almost all samples are strongly enriched in plagioclase relative to the cotectic proportions, an observation consistent with the compositional shift noted in Fig. 1. If the porphyritic magmas were originally related to the aphyric, then enrichment in plagioclase, or alternatively depletion in olivine and clinopyroxene, has led to depletion in MgO relative to the aphyric trend. The fractionation process is of the type termed selective¹², that is, crystals have been removed or added in proportions different from those in which they crystallized.

Densities of aphyric liquids have been calculated using the method of Bottinga and Weill¹³, and, as expected, correlate closely with whole-rock Fe-contents. Thus densities of the liquid fractions of the porphyritic magmas can be estimated if their Fe-contents can be determined.

Direct analysis by microprobe or by phenocryst-groundmass separation has proved impracticable, but estimates have been obtained from modal analyses and phenocryst compositions.

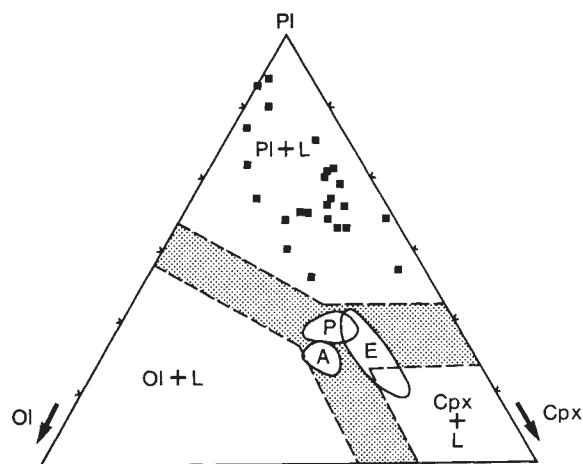


Fig. 2 Plagioclase-rich part of the simple normative basalt tetrahedron at 1 atm (projection from Qz)¹². Shaded zones represent uncertainty in position of loci of liquids in equilibrium with Ol+Pl, Pl+Cpx and Ol+Cpx. Liquids cotectic with regard to Ol+Cpx+Pl lie at the intersection of these zones, in the approximate area of the fields E, A and P. These represent respectively the positions of calculated extracts, aphyric whole-rocks and porphyritic whole-rocks. Filled squares are plotted positions of phenocryst assemblages of porphyritic rocks (in volume %; conversion to weight % makes only a slight difference).

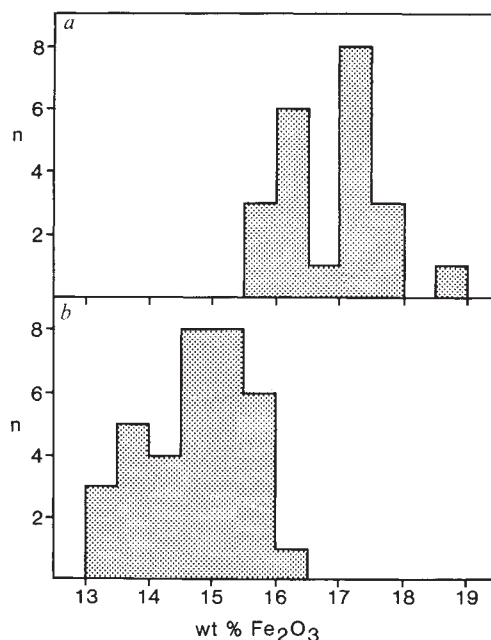


Fig. 3 Histograms for estimated Fe-contents (as total Fe_2O_3) of *a*, porphyritic groundmasses (estimated) and *b*, aphyric rocks (analysed).

Initially calculations were based on the average phenocryst mode for the highly porphyritic rocks (1.77% olivine, 3.35% clinopyroxene, 13.50% plagioclase). Typical Fe-contents for phenocrysts (quoted as Fe_2O_3) from Devey¹¹ are: olivine 32%, clinopyroxene 10%, plagioclase 1.3%, and the average whole-rock Fe_2O_3 is 14.8%. These figures are used to estimate the Fe_2O_3 content of the average groundmass of the highly porphyritic samples as 16.92%. Estimates of the groundmass compositions of 22 individual samples have also been made, using typical Fe-contents for phenocrysts, giving a range of 15.6–18.9% Fe_2O_3 , which overlaps only slightly with the range for the aphyric rocks (Fig. 3).

The porphyritic rocks thus represent magmas that had evolved to lower temperatures than the aphyric. From a consideration of Fe-contents the liquid fractions must have been denser than aphyric magmas, and they carried some of their suspended crystals to the surface, particularly plagioclase. A simple explanation of the porphyritic rocks now emerges, because there is abundant evidence that plagioclase is the one phase under discussion that may become neutrally buoyant in Fe-rich basaltic liquids^{5,13,14}. Plagioclase crystals must have been floated up to the surface with ease, whereas most olivine and clinopyroxene was lost by crystal settling.

This implies strongly that the aphyric magmas have lost all their crystals by the same process, even though the evidence is circumstantial and there is no proof that the aphyric magmas ever actually contained phenocrysts. But we do know that they were undergoing fractional crystallization, and it seems an unnecessarily complex argument to assume that higher-temperature magma batches were crystallizing only by (for example) plating solid phases onto the chamber margins, while lower-temperature batches were producing abundant within-chamber crystals.

We believe that these arguments apply to a wide spectrum of basaltic rocks. Selective fractionation, always involving plagioclase enrichment, has been described from a wide spectrum of basaltic types, including mid-ocean-ridge basalt^{14–19}. The phenomenon is not confined to tholeiites, and may be expected to be operative whenever at least moderately Fe-enriched liquids are produced.

Finally, we seek to reconcile our conclusions with some of those based on fluid dynamics. First, we are not arguing against

convective fractionation, which we cannot detect if its effects are superimposed on those of crystal settling. Second, although we agree with Sparks *et al.*⁶ that turbulent convective motion could suppress crystal settling almost completely, other studies, both theoretical and experimental^{20,21}, have indicated that convection *per se* will not prevent settling. Additionally, if convection in a body ceases or becomes weaker, previously formed crystals can settle out (see ref. 22 for a discussion of 'delayed fractionation').

The yield-strength arguments⁴ still present a puzzle which cannot be answered without further experiments.

In conclusion, we believe that the coincidence we have noted, that is, between the densities of liquids and their capacity to bring crystals to the surface, strongly favours the crystal-settling hypothesis. Whatever other fractionation processes take place within basaltic systems, this simple signal in the erupted record, created by density contrasts between liquids and crystals, comes strongly through. Precisely what fraction of the compositional variation in the suite is generated by this mechanism as opposed to convective fractionation remains unknown.

We thank C. W. Devey for permission to quote his unpublished work. This research was funded by NERC research grant GR3/5350.

Received 12 January; accepted 28 March 1988.

1. Darwin, C. *Geological Observations on the Volcanic Islands, Visited During the Voyage of HMS Beagle* (Smith, Elder, London, 1844).
2. Wager, L. R. & Brown, G. M. *Layered Igneous Rocks* (Oliver & Boyd, Edinburgh, 1968).
3. Campbell, I. H. *Lithos* **11**, 311–323 (1978).
4. McBirney, A. R. & Noyes, R. M. *J. Petr.* **20**, 487–554 (1979).
5. Campbell, I. H., Roeder, P. L. & Dixon, J. M. *Contr. Miner. Petr.* **67**, 369–377 (1978).
6. Sparks, R. S. J., Huppert, H. E. & Turner, J. S. *Phil. Trans. R. Soc. A* **310**, 511–534 (1984).
7. Cox, K. G. & Hawkesworth, C. J. *J. Petr.* **26**, 355–377 (1985).
8. Devey, C. W. & Lightfoot, P. C. *Bull. Volcanol.* **48**, 195–207 (1986).
9. Cox, K. G. & Devey, C. W. *J. Petr.* **28**, 235–238 (1987).
10. Wright, T. L. & Doherty, P. C. *Bull. geol. Soc. Am.* **81**, 1995–2008 (1970).
11. Devey, C. W. thesis, Univ. Oxford (1986).
12. Cox, K. G., Bell, J. D. & Pankhurst, R. J. *The Interpretation of Igneous Rocks* (George Allen & Unwin, London, 1979).
13. Bottinga, Y. & Weill, D. F. *Am. J. Sci.* **272**, 438–475 (1972).
14. Flower, M. F. *Nature* **287**, 530–532 (1980).
15. Gass, I. G., Mallick, D. I. J. & Cox, K. G. *J. geol. Soc. Lond.* **129**, 275–310 (1973).
16. Krishnamurthy, P. & Cox, K. G. *Contr. Miner. Petr.* **73**, 179–189 (1980).
17. Hill, P. G. thesis, Univ. Edinburgh (1974).
18. Armstrong, R. A., Bristow, J. W. & Cox, K. G. *Spec. Publ. geol. Soc. S. Afr.* **13**, 77–86 (1984).
19. Shibata, T., DeLong, S. E. & Walker, D. *Contr. Miner. Petr.* **70**, 89–102 (1979).
20. Marsh, B. D. & Maxey, M. R. *J. Volcan. geotherm. Res.* **24**, 95–105 (1985).
21. Weinstein, S. A., Yuen, D. A. & Olsen, P. L. *Earth planet. Sci. Lett.* **87**, 237–248 (1988).
22. Maaloe, S., Sorensen, I. & Hertogen, J. *J. Petr.* **27**, 439–466 (1986).

Continental rifting at pre-existing lithospheric weaknesses

John A. Dunbar & Dale S. Sawyer*

Department of Geological Sciences, University of Texas at Austin, Austin, Texas 78713, USA

* Institute for Geophysics, University of Texas, 8701 Mopac Boulevard, Austin, Texas 78759, USA

Continental rifting is the poorly understood thermomechanical process by which continents break up and new ocean basins form. Two dominant styles of continental rifting are recognized¹. Crevice style rifts are characterized by initial graben formation, local uplift of the graben shoulders, and little or no volcanism. Arch-volcanic rifts exhibit an initial doming of the crust, extensive volcanism, and late-stage graben formation. The prevailing explanation for the two styles is that crevice rifts form in response to regional horizontal stresses, principally derived from the interaction of the lithospheric plates along their edges, whereas arch-volcanic rifts are thought to be the result of mantle convection currents which upwell directly beneath the rift axis². We propose an alternative working model of continental rifting in which both styles initiate in response to regional horizontal stresses. The different surface expressions of the two styles are explained in terms of differences in the mode of failure at different kinds of pre-existing weaknesses in the continental lithosphere, rather than differences in the nature of the driving forces involved.

Rifts do not occur randomly within continents, but tend to follow orogenic belts, and to avoid or branch around the stronger cratonic regions³⁻⁸. This suggests that the locations of continental rifts are controlled by local variations in the pre-rift strength of continental lithosphere, rather than by the pattern of deep-seated mantle currents⁸. Recent results from rock mechanics indicate that most of the strength of continental lithosphere is contained in two separate strong zones, one in the middle crust and one in the upper mantle⁹ (Fig. 1). Each strong zone is itself a composite of brittle and ductile layers that are tightly bonded at the centre of the strong zone. The total strength of the lithosphere is expected to decrease within orogenic belts as a result of such factors as extensive previous faulting¹⁰, increased volatile content of the mantle¹¹, elevated geotherms¹², increased crustal thickness¹³, and increased lateral heterogeneity within the crust¹⁴. These factors tend to affect the two strong zones differently. For example, a local increase in crustal thickness

from 45 km to 50 km in an orogenic belt would reduce the size of the mantle strong zone without affecting the crustal strong zone (Fig. 1b). Alternatively, an increase in the quartz content of the middle crust, such as would be associated with a granitic arc, would eliminate the centre of the crustal strong zone, but leave the mantle strong zone intact (Fig. 1c). In both of these cases the overall strength of the lithosphere would be reduced by 20%, but each leads to a different mode of continental failure.

We use the forward modelling approach to investigate the mechanical implications of lithospheric failure at different kinds of local weaknesses. Following Chapple¹⁵ and others, we have used Mohr-Coulomb plasticity as a continuum approximation to failure in the brittle regime, whereas deformation in the ductile regime is governed by power-law viscous creep¹⁶. We also assume that the permanent extensional deformation, the flexural response to the permanent extensional deformation, and heat conduction can each be treated separately. The governing equations for the permanent extensional deformation are: (1) the static equilibrium equation for plane strain, (2) either the brittle or ductile constitutive relation, depending on local pressure, temperature, and strain rate conditions, and (3) the compatibility equation for incompressible materials. This set of equations is solved for lithospheric deformation in response to a uniform rate of extension using the finite element method¹⁷. Regional isostatic compensation for this deformation is achieved by allowing the permanently deformed mesh to deform elastically through the separate application of a finite element plane-strain elasticity algorithm¹⁸. Uniform elastic parameters are chosen so that the nominal flexural rigidity of the undeformed lithosphere is 4×10^{23} Nm (ref. 19). The local flexural rigidity during extension decreases with decreasing lithospheric thickness. Thermal effects associated with the thinning of the lithosphere during deformation are accounted for by the application of a finite element heat conduction algorithm²⁰ at the end of each time step.

We consider the extension of an idealized cross-section of continental lithosphere, containing offset crustal and mantle weaknesses of equal magnitude (Fig. 2). At the onset of extension a diffuse zone of simple shear develops in the lower crust between the crustal and mantle weaknesses. At the same time, a pure shear neck forms in the mantle, centred at the mantle weakness. This results in the formation of an asymmetric graben with upturned shoulders over the crustal weakness, and a broad dome over the mantle weakness. Local differential extension between the crust and mantle is taken up along a shear zone in the weak lower crust. As extension proceeds, the deformation near the crustal weakness slows and is progressively shifted to the region

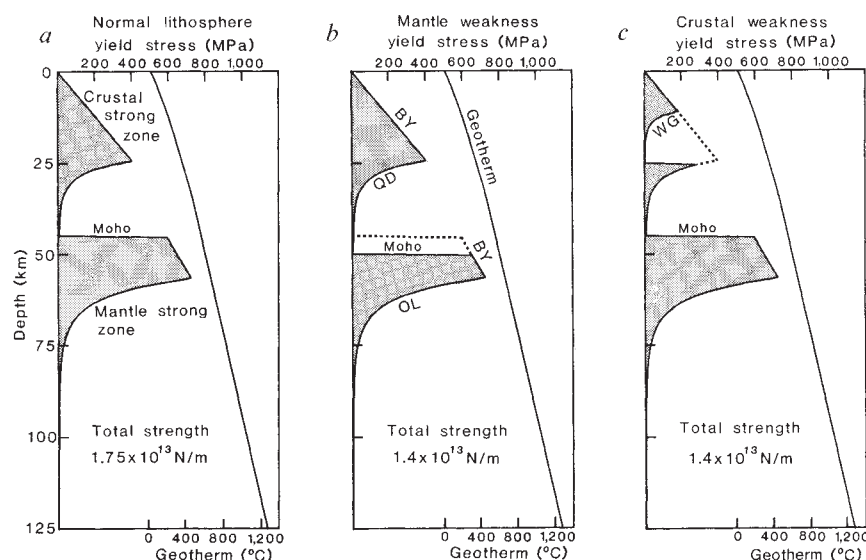


Fig. 1 Yield stress versus depth in continental lithosphere. *a*, Normal orogenic lithosphere with a 45 km-thick crust and strong zones in the middle crust and upper mantle. *b*, Reduction in the size of mantle strong zone associated with an increase in crustal thickness from 45 to 50 km. Dashed lines are the outline from *a*. *c*, Reduction in crustal strong zone associated with a granitic upper and middle crust. Lines labelled BY are from Byerlee's law for frictional sliding in extensional regimes⁹. Lines labelled QD (quartz diorite), OL (olivine), and WG (wet granite) are from ductile flow laws¹⁶ for a strain rate of 10^{-15} s⁻¹ and the geotherm shown. Total strength, in units of force per unit length along the rift axis, is proportional to the area enclosed by the yield stress curve (shaded area).

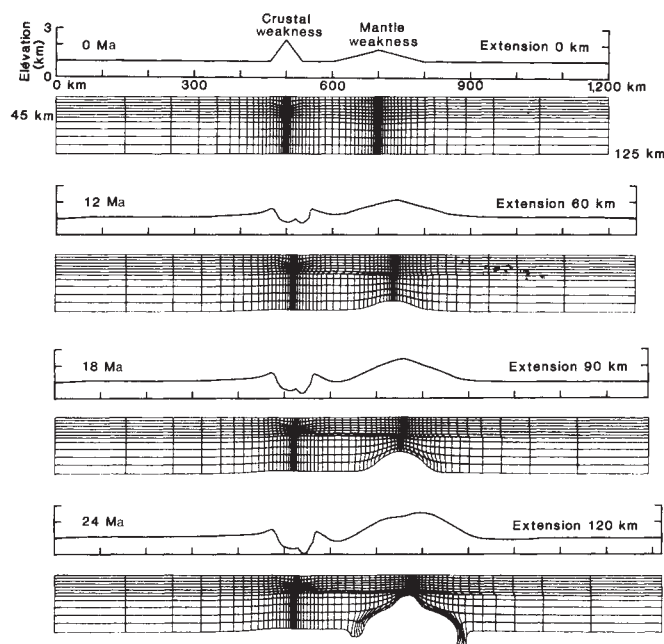


Fig. 2 Deformation during extension of an idealized cross-section of continental lithosphere containing laterally offset crustal and mantle weaknesses. Initial vertical and horizontal lines are boundaries of finite elements which deform with the material. Elements initially range in width from 1 km near the initial weak zones, to 150 km at the edges of the model. Compositions, rheologies, and the initial geotherm are as they appear in Fig. 1, for the normal regions of the model, the section of maximum crustal weakness, and the section of maximum mantle weakness. Initial topography reflects the local isostatic compensation for mass deficiencies associated with the granitic inclusion in the crust and the region of thick crust. The left boundary is held fixed, whereas the right boundary is extended at a constant rate of 0.5 cm per year, in constant time steps of 0.5 Myr.

near the mantle weakness (Fig. 3). Also, the slope of the lithosphere-asthenosphere boundary on the flanks of the necked region increases. Late in the model run, the increasing slope coupled with the inherent gravitational instability of the lower lithosphere²¹ causes a flow of material at the base of the lithosphere away from the axis of mantle extension. By the end of the model run, this material collects in 20 km-wide drips at the base of the lithosphere and the deformation within the crust becomes focused over the region of mantle failure.

Differences in the pattern of deformation associated with failure at crustal and mantle weaknesses result in marked differences in the pressure-temperature regime of the lithosphere and upper asthenosphere near the two types of weakness. The top of the asthenosphere near the crustal weakness rises only slightly during the model run and never approaches the expected onset of melt segregation at 2% partial melt²² (Fig. 4). Little or no volcanism would be expected in this region. In contrast, the top of the asthenosphere near the mantle weakness reaches the 2% partial melt condition shortly after 15 Myr of extension, and approaches 10% partial melt by the end of the model run. The rapid rise of the base of the lithosphere in this region also causes the progressive heating of the lower crust, where low pressure fractionation could produce volumes of light salic magmas²³. This may provide an explanation for the evolution of magma from basic to more salic compositions with time, observed in some arch-volcanic rifts^{23,24}. The surface expression of rifting predicted by the model would then be an initial formation of a graben with upturned shoulders and little or no volcanism over

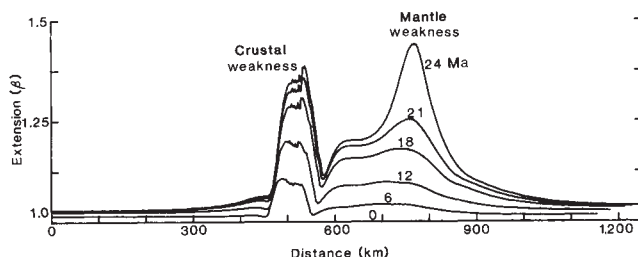


Fig. 3 Extension at the top of the crust. Model extension is measured as the ratio of current length of an element side over the original length at various times in the model run. This extension would manifest itself as surface faulting. During the early stages, extension increases more rapidly at the crustal weakness than at the mantle weakness. During the late stages the pattern reverses.

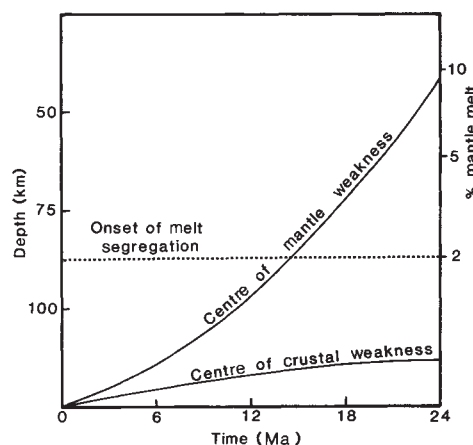


Fig. 4 Degree of partial melting at the top of the asthenosphere. The top of the asthenosphere, initially at 1,280 °C, cools adiabatically and undergoes partial melting as it rises to lower pressure regimes. The onset of melt segregation and melt per cent versus depth are after Ringwood²².

the failure at the crustal weakness. Over the failure at the mantle weakness, the model predicts initial doming, extensive volcanism, and late-stage graben development.

The Sierra Nevada uplift and Great Basin depression of Central California may represent a geological example, in the early stages of development, of offset crust and mantle failure²⁵. Teleseismic and seismic refraction data suggest that the mantle is currently failing along an axis of thick crust beneath the Sierra Nevada, resulting in the uplift of the Sierra Nevada and lithospheric scale drips which extend 175 km or more into the asthenosphere^{25,26}. Geological evidence indicates that the crust within the Great Basin, some 200 km to the east, is undergoing extension along a combination of westward dipping low-angle and high-angle normal faults²⁷.

In this paper we describe a case in which crustal and mantle weaknesses of equal magnitude are laterally offset by 200 km. In general, the weaknesses could be separated by a different distance or superimposed. Also, either weakness could be less well developed or absent all together. A continuum of possible pre-rift configurations exists in which one endmember is represented by an isolated crustal weakness and the other by an isolated mantle weakness. Failure at isolated crustal weaknesses begins with graben formation, but because mantle extension is everywhere diffuse, the onset of volcanism is delayed until the late stages of continental breakup. In contrast, failure at an isolated mantle weakness begins with the broad subsidence of the crust, followed closely by volcanism along the prerift axis and uplift of the rift shoulders. The range of possible rifting

styles between these two extremes may provide explanations for such diverse rifting phenomena as crustal doming, the occurrence of so called 'wet' rifts (with volcanism), and 'dry' rifts (without volcanism)²³, uplifted graben shoulders²¹, apparent depth-dependent extension²⁸, and the apparent simple shear²⁹ or delamination of the lithosphere³⁰. This alternative working model of continental rifting is consistent with current ideas about both the rheology of the lithosphere and the driving forces of plate tectonics.

We thank Roger Buck for constructive reviews of the manuscript.

Received 1 February; accepted 11 April 1988.

1. Milanovsky, E. E. *Tectonophysics* **15**, 65-70 (1972).
2. Sengör, A. M. & Burke, K. *Geophys. Res. Lett.* **5**, 419-421 (1978).
3. Wilson, J. T. *Nature* **207**, 343-347 (1965).
4. McConnell, R. B. *Bull. geol. Soc. Am.* **83**, 2549-2572 (1972).
5. King, B. C. in *Tectonics and Geophysics of Continental Rifts* (eds Ramberg, I. B. & Neumann, E. R.) 347-350 (Reidel, Dordrecht, 1978).
6. Fuchs, K. in *Approaches to Taphrogenesis* (eds Illies, J. H. & Fuchs, K.) 420-432 (Schweizerbart, Stuttgart, 1974).
7. Illies, J. H. in *Tectonics and Geophysics of Continental Rifts* (eds Ramberg, I. B. & Neumann, E. R.) 63-71 (Reidel, Dordrecht, 1978).
8. Sykes, L. R. *Rev. Geophys. Space Phys.* **16**, 621-688 (1978).
9. Brace, W. F. & Kohlstedt, D. L. *J. geophys. Res.* **85**, 6248-6252 (1980).
10. Meissner, R. *The Continental Crust*, 162-167 (Academic, New York, 1986).
11. Pollack, H. N. *Earth planet. Sci. Lett.* **80**, 175-182 (1986).
12. Ballard, S. & Pollack, H. N. *Earth planet. Sci. Lett.* **85**, 253-264 (1987).
13. Vink, G. E., Morgan, W. J. & Zhao, W. J. *geophys. Res.* **89**, 10072-10076 (1984).
14. Dewey, J. F. in *The Nature of the Lower Continental Crust* (eds Dawson, J. B., Carswell, D. A., Hall, J. & Wedepohl, K. H.) 71-78 (Geological Society, London, 1986).
15. Chapple, W. M. *Bull. geol. Soc. Am.* **89**, 1189-1198 (1978).
16. Kirby, S. H. *Rev. Geophys. Space Phys.* **21**, 1458-1487 (1983).
17. Zienkiewicz, O. C. & Godbole, P. N. in *Finite Elements in Fluids* Vol. 1, Ch. 2 (eds Gallagher, R. H., Oden, J. T., Taylor, C. & Zienkiewicz, O. C.) (Wiley, New York, 1975).
18. Zienkiewicz, O. C. *The Finite Element Method* 93-118 (McGraw-Hill, London, 1977).
19. Walcott, A. B. *J. geophys. Res.* **75**, 3941-3954 (1970).
20. Huebner, K. H. & Thornton, E. A. *The Finite Element Method for Engineers* 423-427 (Wiley, New York, 1982).
21. Buck, W. R. *Earth planet. Sci. Lett.* **77**, 362-372 (1986).
22. Ringwood, A. E. *Composition and Petrology of the Earth's Mantle* 150-169 (McGraw-Hill, New York, 1975).
23. Barberi, F., Santacrose, R. & Varet, J. in *Continental and Oceanic Rifts, Geodyn. Ser. Vol. 8* (ed. Palmason, G.) 293-309 (AGU, Washington, D.C., 1982).
24. Mohr, P. *Trans. Am. geophys. Un.* **68**, 721-730 (1987).
25. Jones, C. H. *Tectonics* **6**, 449-473 (1987).
26. Kerr, R. A. *Science* **239**, 978-979 (1988).
27. Burchfiel, B. C., Hodges, K. V. & Royden, L. H. *J. geophys. Res.* **92**, 10422-10426 (1987).
28. Rowley, D. B. & Sahagian, D. *Geology* **14**, 32-35 (1986).
29. Wernicke, B. *Can. J. Earth Sci.* **22**, 108-125 (1984).
30. Lister, G. S., Etheridge, M. A. & Symonds, P. A. *Geology* **14**, 246-250 (1986).

Network model of shape-from-shading: neural function arises from both receptive and projective fields

Sidney R. Lehky & Terrence J. Sejnowski

Department of Biophysics, Johns Hopkins University, Baltimore, Maryland 21218, USA

It is not known how the visual system is organized to extract information about shape from the continuous gradations of light and dark found on shaded surfaces of three-dimensional objects^{1,2}. To investigate this question^{3,4}, we used a learning algorithm to construct a neural network model which determines surface curvatures from images of simple geometrical surfaces. The receptive fields developed by units in the network were surprisingly similar to the actual receptive fields of neurons observed in the visual cortex^{5,6} which are commonly believed to be 'edge' or 'bar' detectors, but have never previously been associated with shading. Thus, our study illustrates the difficulty of trying to deduce neuronal function solely from determination of their receptive fields. It is also important to consider the connections a neuron makes with other neurons in subsequent stages of processing, which we call its 'projective field'.

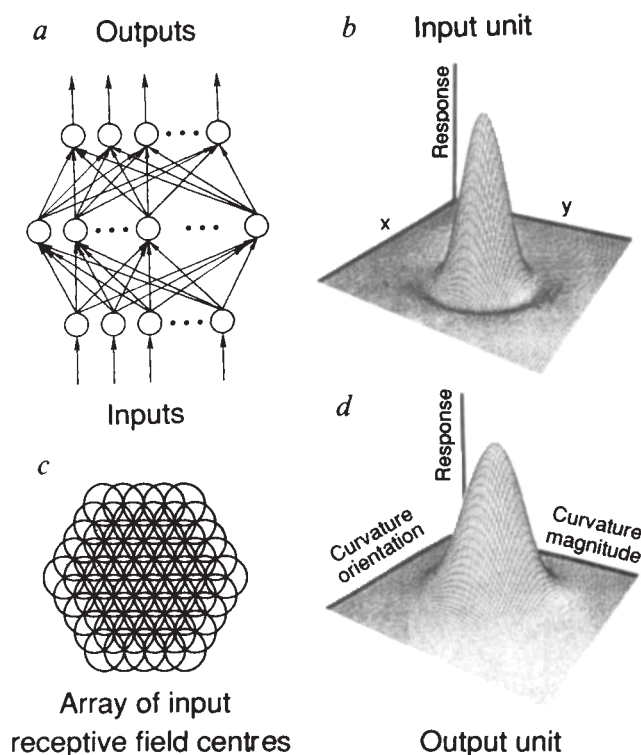


Fig. 1 Organization of neural network that extracts surface curvatures from images of shaded surfaces. *a*, Diagram of three-layer network. Each unit projects to all units in the subsequent layer. The responses of the units in the input layer are determined by the environment. The responses of each unit in the hidden and output layers are determined by summing the activities from all units in preceding layer, weighted by connection strengths which can be positive or negative, and then passed through a sigmoid nonlinearity. Unit activities can assume any value between 0.0-1.0. *b*, Input-unit receptive field, formed by the Laplacian of a two-dimensional gaussian. On-centre units have an excitatory centre and an inhibitory surround, while off-centre units have opposite centre/surround polarities. The on- and off-centre terminology does not imply any temporal properties for these units. *c*, Receptive field centres of input units overlapped in a hexagonal array. Images were sampled by both on-centre and off-centre arrays, which were spatially superimposed. *d*, Output-unit response curve, tuned to both curvature magnitude and orientation. The maximum response of each output unit was produced by a different combination of those two curvature parameters. The magnitude axis is on logarithmic scale. Multi-dimensional responses such as this are common in the visual cortex for various parameters, although units selective for surface curvature have not been reported.

The specific task we set for the network was to determine the magnitudes and orientations of the two principal surface curvatures at the centre of each input surface, and to do this independently both of lateral translations of the surface within a small patch of the visual field, and also independently of the direction of illumination. Surface curvature depends upon the direction of travel along a surface. The principle curvatures are the maximum and minimum curvatures for all trajectories through a particular point, which are always perpendicular to each other, and are good descriptors of local shape.

The network had three layers (Fig. 1*a*): an input layer (122 units), an output layer (24 units), and an intermediate hidden layer (27 units). The input layer consist of arrays of units with circular receptive fields (Fig. 1*b, c*), similar to neurons found in the retina and the lateral geniculate nucleus. Output units were selective for both the magnitude and orientation of curvature (Fig. 2*d*). Because of its non-monotonic, tuned response, the activity of a single output unit represented the curvature

Fig. 2 Typical input image and resulting activity levels within a trained network. *a*, Example of an elliptical paraboloid surface. (The flat base did not fall within the input field of the network). *b*, One of 2,000 images used to train the network, synthesized by calculating light reflected from the paraboloid surface. Each image differed in the magnitudes and orientation of the two principal curvatures, in the slant and tilt of illumination, and in the location of the surface centre within the input field. All image parameters were randomly selected from a uniform distribution. The curvature magnitude ranged from 2 deg^{-1} to 32 deg^{-1} and also -2 deg^{-1} to -32 deg^{-1} , and the curvature orientation from 0° to 180° . The centre of the paraboloid could fall anywhere within the central third of the input field, and the surface normal at the paraboloid centre was always perpendicular to the image plane. Surface reflection was Lambertian, or matte. Illumination came predominantly from one direction but was partially diffused to eliminate sharp shadow edges. The illumination slant fell between 0° – 60° . The network was trained to interpret images assuming that illumination came from above (tilt between 0° – 180°), and that the signs of both curvatures were the same (that is, the surface was convex or concave). *c*, The network response to an image. The area of a black square indicates a unit's activity. Double hexagons show the responses of 61 on-centre and 61 off-centre input units, calculated by convolving their receptive fields with the image. The responses were rectified, and so they only assumed positive values. These input units cause activity in the 27 hidden units, arranged in a 3×9 array above the hexagons. The hidden units in turn project to the output layer of 24 units, shown in a 4×6 array. This output should be compared with the other 4×6 array at the very top, which shows a correct response to the image. Units within a 4×6 array are arranged as follows. The six columns correspond to different peaks in orientation tuning, at 0° , 30° , 60° , 90° , 120° and 150° . The rows correspond to different curvature magnitudes: the top two rows code for positive (tuning peak: $+8 \text{ deg}^{-1}$) and negative (tuning peak: -8 deg^{-1}) magnitudes of the smaller of the two principal curvatures (C_s), while the bottom two rows code the same for the larger principal curvature (C_L) (same tuning peaks). Curvature orientation is unambiguously coded by the pattern of activity in six overlapping orientation-tuning curves. Representation of curvature magnitude, however, remains degenerate because output unit tuning curves in that domain do not overlap. An output can therefore correspond to two curvature magnitudes, which the network cannot distinguish. This remaining ambiguity could be resolved with a larger network containing units that are responsive at different spatial scales.

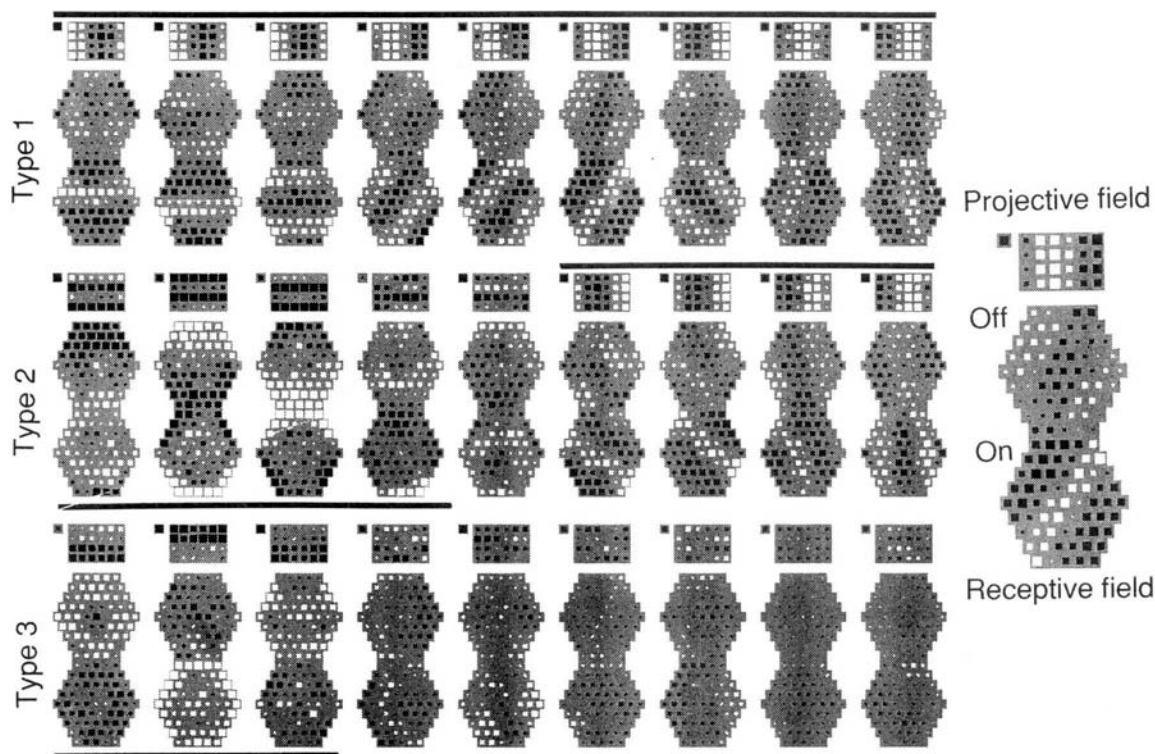
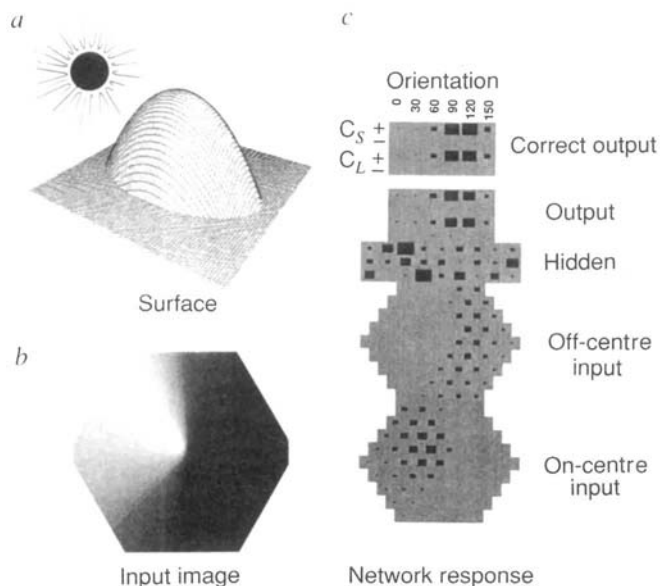


Fig. 3 Connection strength in a typical network. Excitatory weights are white and inhibitory ones are black, and the areas of the squares indicate the connection strengths. Each hidden unit is represented by one hourglass-shaped icon, showing its receptive field (double hexagons) and projective field (4×6 array at the top). The organization of units in the 4×6 array is as described in Fig. 2c. The isolated square at the left of each icon indicates the unit's bias (equivalent to a negative threshold). Black horizontal lines group units that have the same type of projective field organization

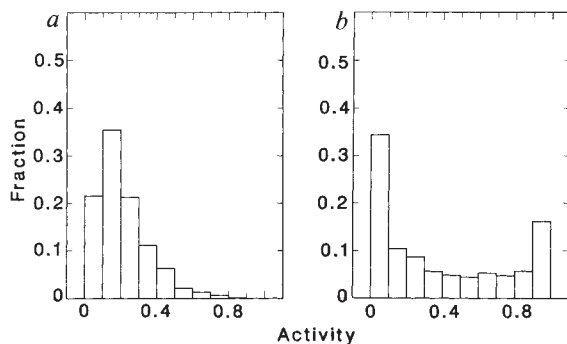


Fig. 4 Distribution of activity levels for two hidden units when the network was presented with the 2,000 images described in Fig. 2b. For each image, units gave responses between 0.0 and 1.0, which were grouped into ten bins. *a*, Histogram with a unimodal distribution typical of orientation-selective units (type 1) and of units selective for the relative magnitudes of the principal curvatures (type 3). These units tend to be activated over a range of intermediate levels when presented with many inputs, and appear to act as continuous, tuned filters indicating the values of their respective parameters. *b*, Histogram with typical bimodal distribution for a unit discriminating between positive and negative curvatures (type 2). These units are like feature detectors, tending to be either fully on or off to indicate whether a surface is convex or concave.

parameters in a degenerate manner; that is, various input images could lead to the same response. To resolve this ambiguity, curvature parameters were encoded by the pattern of activity in a population of output units having different, but overlapping tuning curves, analogous to the way that colour can be encoded by the pattern of activity in three broadly tuned channels. This network was intended to model processing for only a small patch of the visual field, about the size handled by a single cortical column. It would have to be replicated at different locations to cover the entire field, perhaps with all components feeding into a higher level network to integrate the local analyses.

Given this three-layer network architecture, the 'back-propagation' learning algorithm⁷ was used to organize the properties of the hidden units to provide a transform between the retinotopic space of the input units and the two-dimensional magnitude and orientation parameter space of the outputs. Images of elliptic paraboloid surfaces were used as inputs (Fig. 2a, b). Sharp edges were excluded from the images, and so the only cues available for computing curvatures were in the shading. The network was presented with many images, and, for each input, responses were propagated up to the output units. The actual output was then compared with the correct output for that image, and all connection strengths in the network were slightly modified to reduce error in the manner specified by the algorithm. Gradually, the initially random connection strengths became organized. The correlation between the correct and the actual outputs reached a plateau of 0.88 after 40,000 presentations, and the network generalized well for images that were not part of the training set. Increasing the number of hidden units failed to improve the network performance, although it did deteriorate when there were too few hidden units. No biological significance is claimed for the algorithm by which the network developed but, rather, the focus of interest is on the resulting mature network.

An example of the network's response to an image is given in Fig. 2c and the network connection strengths underlying this response are shown in Fig. 3. Each of the 27 hourglass-shaped icons represents the connections associated with one hidden unit. The double hexagons in each icon show the connections from all input units to that hidden unit (that is, the receptive field), and the 4×6 array at the top shows the connections between that hidden unit and all output units (that is, the projective field). Repeating the learning procedure starting from completely different sets of random weights resulted in essentially the same pattern of connections.

The receptive fields in Fig. 3 are reminiscent of those in the visual cortex^{5,6,8}. Excitatory and inhibitory connections are often organized in an orientation-specific and multi-lobed manner, although some are more or less circularly-symmetrical. Upon examining projective fields, however, three types become apparent: type 1 has a vertical pattern of organization to the 4×6 array of weights; type 2 has a horizontal organization with alternate rows being similar; and type 3 has a horizontal organization with adjacent rows being similar. These classes of hidden units appear to provide information to output units about, respectively, the orientation of the principal curvatures (type 1), their signs (convexity/concavity) (type 2), and their relative magnitudes (type 3). The units had different response distributions when presented with many stimuli. Type 1 and type 3 had unimodal distributions (Fig. 4a), whereas type 2 units had bimodal distributions (Fig. 4b). Based on these distributions, we interpret types 1 and 3 as being filters that indicate values for their respective parameters, and type 2 units as being feature detectors that discriminate between discrete alternatives (convexity and concavity). A few hidden units were difficult to classify, and four failed to develop large weights.

We tried probing the units with simulated bars of light and found that the responses of the hidden units were easily predictable from the pattern of excitatory and inhibitory connections they received from the input units, and that most of these responses appear similar to those of simple cells in the visual cortex^{5,8}. In contrast, it required extensive trial and error to find the optimal stimulus for the output units, but this was not surprising as each output unit received convergent inputs from all 27 hidden-unit receptive fields. Some output units had strong 'end-stopped inhibition', similar to that of some complex cells in the cortex⁶. In these units, responses dropped precipitously when the bar length was extended beyond a certain point.

Examination of the receptive fields of individual units does not make apparent what the network is doing, and interpretations other than that of extracting curvatures from shaded images are likely to spring to mind. While this model network obviously does not establish that receptive fields in the cortex which resemble those developed by the network are engaged in shading analysis, it does raise questions about conventional interpretations of the functions of receptive fields, not only in visual pathways, but in other sensory systems as well. Understanding the function of a neuron within a network appears to require not only knowledge of the pattern of input connections forming its receptive field, but also knowledge of the pattern of output connections, which forms its projective field. Indeed, the same neuron may have a number of different functions if it projects to several regions.

This work was supported by an NSF Presidential Young Investigator Award to TJS and a Sloan Foundation grant to TJS and Dr G. F. Poggio.

Received 7 March; accepted 25 April 1988.

1. Ramachandran, V. S. *Nature* **331**, 163-166 (1988).
2. Mingolla, E. & Todd, J. T. *Biol. Cyber.* **53**, 137-151 (1986).
3. Ikeuchi, K. & Horn, B. K. P. *Art. Intell.* **17**, 141-184 (1981).
4. Pentland, A. P. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 170-187 (1984).

5. Hubel, D. H. & Wiesel, T. N. *J. Physiol., Lond.* **160**, 106-154 (1962).
6. Hubel, D. H. & Wiesel, T. N. *J. Neurophysiol.* **28**, 229-289 (1965).
7. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1 (eds Rumelhart, D. E. & McClelland, J. L.) 318-362 (MIT Press, Cambridge, 1986).
8. Mullikan, W. H., Jones, J. P. & Palmer, L. A. *J. Neurophysiol.* **52**, 372-387 (1984).

Epstein-Barr virus genome-positive T lymphocytes in a boy with chronic active EBV infection associated with Kawasaki-like disease

Hideaki Kikuta, Yuichi Taguchi, Kazuhiro Tomizawa, Kimikazu Kojima, Nobuaki Kawamura, Akihito Ishizaka, Yukio Sakiyama, Shuzo Matsumoto, Shosuke Imai†, Toshifumi Kinoshita†, Shigeki Koizumi†, Toyoro Osato†, Ichiro Kobayashi‡, Isamu Hamada‡ & Kanji Hirai§

Department of Pediatrics, Hokkaido University School of Medicine, Sapporo 060, Japan

† Department of Virology, Cancer Institute, Hokkaido University School of Medicine, Sapporo 060, Japan

‡ Pediatric Clinic, National Hakodate Hospital, Hakodate 041, Japan

§ Department of Molecular Biology, Tokai University School of Medicine, Isehara 259-11, Japan

Epstein-Barr virus (EBV), a ubiquitous human herpesvirus and an aetiological agent of infectious mononucleosis, has a unique tropism for B lymphocytes¹. Clinical and laboratory features of chronic active EBV infections are chronic or persistent infectious mononucleosis-like symptoms and high antibody titre against early antigens (EA)²⁻⁴. Kawasaki disease (KD), aetiology unknown, is thought to be self-limited immunologically mediated vasculitis. Clinical features of KD are fever, rash, mucositis, lymphadenopathy and coronary artery damage^{5,6}. We report here a child with chronic active EBV infection accompanied by dilatation of coronary arteries. All the EBV-determined nuclear antigen (EBNA)-positive lymphocytes had exclusively CD4 antigen, as revealed by dual staining immunofluorescence analysis. Southern blot hybridization showed that the purified CD4+ cells harboured EBV genome.

Our subject, a two-year-old boy with no family history of disease, was admitted to hospital in August 1987. He had a three-month history of continued spiking fever, hepatosplenomegaly, generalized lymphadenopathy and dilatation of

Table 1 Phenotypic distribution of the patient's peripheral blood lymphocytes

Reagent	Percentage of positive cells
T11 (CD2)	85
T3 (CD3)	79
T4 (CD4)	72
T8 (CD8)	8
WT31 (Ti)	91
2H4	65
NKH-1	4
B2 (CD21)	8
B4 (CD19)	8
anti-human Ig	10
anti-FcεR (CD23)	7
IL-2R	1
I2 (Ia/DR)	52
T3(CD3) + I2(Ia/DR)	59*
T4(CD4) + I2(Ia/DR)	54*
T8(CD8) + I2(Ia/DR)	4*

Analysis of the patient's peripheral blood mononuclear cell surface phenotypes was done by flow cytometry with Coulter Co Epics-C flow cytometer after staining with phycoerythrin (PE)- or fluorescein-isothiocyanate (FITC)-conjugated monoclonal antibodies (Coulter Clone), or FITC-conjugated rabbit anti-human immunoglobulins (DAKOPATTS) by direct immunofluorescence.

* Percentage of double positive cells by dual staining.

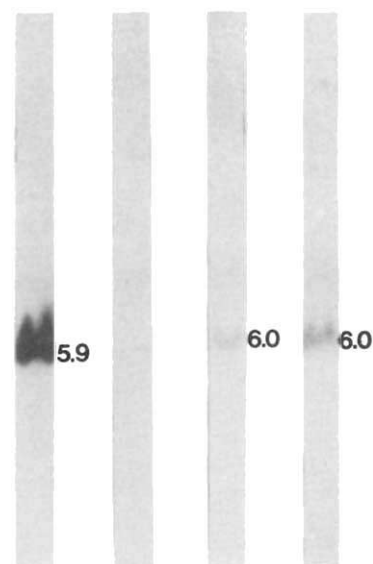


Fig. 1 Detection of EBV DNA in the peripheral blood mononuclear cells and the sorted CD4+ cells by Southern hybridization. **Methods.** DNA was extracted from the peripheral blood mononuclear cells, the sorted CD4+ cells from the peripheral blood mononuclear cells (PBMCs), B95-8 cells and BJAB cells. DNA was cleaved with *Bam*H1. The DNA was subjected to electrophoresis through 0.6% agarose gel and transferred on nitrocellulose membrane filter. Each filter was hybridized with ³²P-labelled cloned fragment *Bam*H1-H2 of EBV DNA for 48 h at 41 °C in 1 × SSC, 50% formamide, 0.5% sodium dodecyl sulphate (SDS) and heat-denatured salmon sperm DNA (100 µg ml⁻¹). After hybridization, the filters were washed three times at room temperature in 0.1 × SSC, 0.1% SDS, then incubated for 1 h at 50 °C three times. The filters were then dried and exposed to a Sakura X-ray film at -80 °C. Lane 1 contains 3 µg of B95-8 cell DNA. Lane 2 contains 5 µg of BJAB cell DNA. Lane 3 contains 5 µg of the PBMC DNA. Lane 4 contains 2 µg of the sorted CD4+ cell DNA. The size of each fragment is indicated in kb on the right of corresponding bands.

the coronary arteries but no history of KD until March 1987, when a cervical lymphadenopathy became evident. At that time, the serological tests specific for EBV demonstrated a viral capsid antigen (VCA) IgG antibody titre of 1:320 and an EBNA antibody titre of <1:5, findings indicative of a recent EBV infection. In May 1987, dilatations of coronary arteries developed in addition to high fever and hepatosplenomegaly.

On admission, the following laboratory results were obtained: haemoglobin, 8.4 g 100 ml⁻¹; haematocrit, 26.4%; white-cell count, 8,900 with 47% neutrophils, 48% lymphocytes (0% atypical lymphocytes) and 5% monocytes; and, platelet count, 264,000. Other serum values included the serum glutamic oxaloacetic transaminase (321 IU l⁻¹), the serum glutamic pyruvic transaminase (747 IU l⁻¹), the lactate dehydrogenase (1,518 IU l⁻¹) and the total protein (6.3 g 100 ml⁻¹). Quantitation of serum immunoglobulins was as follows: IgG, 1,100 mg 100 ml⁻¹; IgM, 134 mg 100 ml⁻¹; and, IgA, 79 mg 100 ml⁻¹. The natural killer-cell activity was normal as was the production of α and γ interferon and the production of interleukin-2.

Flow cytometry with a Coulter Co Epics-C flow cytometer revealed an increased population of CD4+ cells, a decreased population of CD8+ cells and a high Ia/diffuse (D) and restricted (R) expression on the CD4+ cells (Table 1). A lymph-node biopsy specimen showed non-neoplastic lymphoid hyperplasia and an examination of the bone marrow showed no leukemic cells. Echocardiography revealed that the right and left coronary

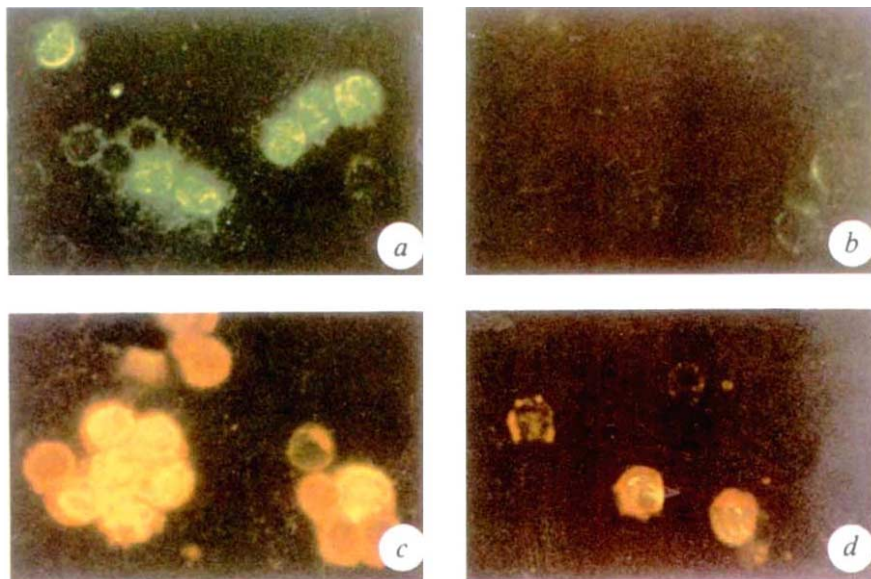


Fig. 2 Dual staining immunofluorescence analysis of EBNA-positive cells in the patient's peripheral blood mononuclear cells. *a*, EBNA staining by seropositive reference serum. *b*, EBNA staining by seronegative reference serum. *c*, Dual staining of EBNA and CD4. *d*, Dual staining of EBNA and Ia/DR.

Methods. Acetone:methanol (1:1)-fixed cell smears were incubated with seropositive reference serum (*a, c, d*) or seronegative reference serum (VCA-IgG titre of <1:5, EA-IgG titre of <1:5, EBNA titre of <1:5) (*b*) at 37 °C for 45 minutes. Thereafter the preparations were washed with PBS. Subsequently the preparations were reacted with FITC-conjugated goat anti-human C3c (DAKOPATTS) at 37 °C for 30 minutes. After additional washing with PBS, they were incubated with T4 (*c*) or I2 (*d*) monoclonal antibody (Coulter Clone) at 37 °C for 45 minutes. After another washing with PBS, the preparations were incubated with tetramethylrhodamine isothiocyanates-conjugated rabbit anti-mouse IgG (DAKOPATTS).

arteries were 3 mm and 6 mm in diameter, respectively. The patient had an IgG antibody titre to VCA of 1:10,240, an IgM antibody titre to VCA of <1:5, an IgG antibody titre to EA-DR of 1:2,560 and an IgG antibody titre to EBNA of 1:160. There were no elevations in antibody titres to herpesvirus type-1 and type-2, cytomegalovirus, measles virus, rubella virus or adenovirus type-5. Antibodies to human T-lymphotropic virus (HTLV-1) and human immunodeficiency virus (HIV) were not detectable. There were 36% EBNA-positive cells in the peripheral blood mononuclear cells, 5% in the aspiration specimen of bone marrow and 7% in the mononuclear cells of cervical lymph node. EBV genomes were also detected in the peripheral blood mononuclear cells using Southern blot hybridization with the ³²P-*Bam*HI-H2 fragment of the *EBNA-2* gene of the EBV M-ABA strain (Fig. 1). The *Bam*HI-H fragment was used to differentiate two types, A and B, of EBV isolates by Zimmer *et al.*⁷. Because the *Bam*HI-H2 hybridized to the 6 kilobase (kb) fragment of the peripheral blood mononuclear cells as well as to the *Bam*HI-H (5.9 kb) of the B95-8 strain, the EBV in the peripheral blood mononuclear cells was shown to be type A, generally observed in the northern hemisphere. Because only B lymphocytes are infected by EBV¹, it is striking that such a high proportion (36%) of EBNA-positive cells was present in the peripheral blood mononuclear cells.

To characterize the EBNA-positive cells⁸, we used dual-staining immunofluorescence analysis with seropositive reference serum and a wide variety of monoclonal antibodies. This defined the phenotype of EBNA-positive cells to be CD2, CD3, CD4 and Ia/DR positive, and CD8, CD19, CD20, CD21, 2H4 and NKH-1 negative (Table 2). The vast majority of EBNA-positive cells from the peripheral lymphocytes displayed features of activated CD4+ helper-inducer lymphocytes. B-lymphocyte surface markers were not expressed on the EBNA-positive cells (Fig. 2). To confirm which cell population was responsible for the presence of EBNA and EBV genomes, we sorted CD4+ cells on a fluorescence-activated cell sorter. Fifty-three per cent of the sorted CD4+ cells (99.2% purified) were EBNA-positive by anti-complement immunofluorescence and the sorted CD4+ cells harboured EBV genomes, as determined by Southern blot analysis using ³²P-*Bam*HI-H2 (Fig. 1).

Spontaneous lymphoblastoid cell lines could not be established from the peripheral blood mononuclear cells, despite several attempts. EBNA-positive T lymphocytes did not express EA⁹ or VCA¹⁰ in culture, thereby indicating that EBV may infect

T lymphocytes abortively *in vitro*. Other viruses inducing cytopathic effects, such as human herpesvirus-6 (HHV-6)^{11,12}, were never isolated from the peripheral blood mononuclear cells.

The phenotype of EBV-infected cells in the peripheral blood mononuclear cells was that of activated CD4+ helper-inducer lymphocytes, and CD21 antigen, the EBV receptor¹³, was not expressed on the surface of EBNA-positive cells. Why EBV-receptor-negative T lymphocytes of the patient were infected with EBV was not answered. We previously reported that KD appears to be closely associated with an unusual primary infection with EBV¹⁴, and that EBV events were suppressed in KD¹⁴⁻¹⁶. In contrast, EBV events were persistently activated in chronic EBV infection which showed a mirror image to that in KD. EBV may play some role in the development of coronary artery damage in chronic EBV infection and KD.

Table 2 Percentage of EBNA-positive cells in various surface antigen-positive lymphocytes

Surface antigen	Percentage of EBNA-positive cells
CD2+	43
CD3+	40
CD4+	46
CD8+	0
2H4+	0
NKH-1+	0
CD19+	0
CD20+	0
CD21+	0
Ia/DR+	55

Patient's peripheral blood mononuclear cells were separated on Ficoll-Conray gradients. The cell smears were fixed for 1 min in ice-cold acetone:methanol (1:1) and stained for EBNA according to anti-complement immunofluorescence method of Reedman and Klein⁸. The cell smears were incubated with seropositive reference serum (VCA-IgG titre of 1:640; EA-IgG titre of 1:40; EBNA titre of 1:640) containing complement at 37 °C for 45 min. After being washed with phosphate buffered saline (PBS), FITC-conjugated goat anti-human C3c (DAKOPATTS) were added. The preparations were maintained at 37 °C for 30 min. After staining EBNA, the preparations were stained with PE-conjugated monoclonal antibodies (Coulter Clone) by direct immunofluorescence. All the stained preparations were examined under a Nikon Universal Research Microscope, Microphoto FX with Epi-Fluorescence equipment. The proportion of EBNA positive cells in the peripheral blood mononuclear cells was 36%.

We thank Dr Georg W. Bornkamm for providing pM-BamH2, Dr M. Suemura for providing PE-conjugated anti-Fc ϵ R monoclonal antibody, Becton Dickinson Co. for providing FITC-conjugated WT31 monoclonal antibody, and M. Ohara for comments. This work was supported by the Ministry of Education, Science and Culture, and the Ministry of Health and Welfare for Primary Immunodeficiency Diseases and Comprehensive 10-year Strategy for Cancer Control, Japan.

Received 18 February; accepted 18 April 1988.

1. Epstein, M. A. & Achong, B. G. The Epstein-Barr Virus (eds Epstein, M. A. & Achong, B. G.) 1-22 (Springer, Berlin, 1979).
2. Alfieri, C., Ghibu, F. & Joncus, J. H. *Can. med. Assoc. J.* **131**, 1249-1252 (1984).
3. Jones, J. F. *et al. Ann. Int. Med.* **102**, 1-7 (1985).
4. Alfieri, C. & Joncus, J. H. *J. Virol.* **61**, 3306-3309 (1987).
5. Kawasaki, T. *et al. Pediatrics* **54**, 271-276 (1974).
6. Fujiwara, H. & Hamashima, Y. *Pediatrics* **61**, 100-107 (1978).
7. Zimmer, U. *et al. Virology* **154**, 56-66 (1986).
8. Reedman, B. M. & Klein, G. *Int. J. Cancer* **11**, 499-520 (1973).
9. Henle, W. *et al. Science* **169**, 188-190 (1970).
10. Henle, G. & Henle, W. *J. Bacteriol.* **91**, 1248-1256 (1966).
11. Salahuddin, B. Z. *et al. Science* **234**, 596-601 (1986).
12. Josephs, S. F. *et al. Science* **234**, 601-603 (1986).
13. Tanner, J., Weis, J., Fearon, D., Whang, Y. & Kieff, E. *Cell* **50**, 203-213 (1987).
14. Kikuta, H. *et al. Pediatrics* **73**, 413-414 (1984).
15. Iwanaga, M. *et al. Lancet* **i**, 938-939 (1981).
16. Kikuta, H. *et al. Acta Paediatr. Jpn* **28**, 120 (1986).

Sequence of simian immunodeficiency virus from African green monkey, a new member of the HIV/SIV group

Masashi Fukasawa*, Tomoyuki Miura*, Akira Hasegawa†, Shigeru Morikawa‡, Hajime Tsujimoto*, Keizaburo Miki†, Takashi Kitamura‡ & Masanori Hayami*

* Department of Animal Pathology, Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo 108, Japan

† Toa Nenryo Kogyo K.K., Ohi-machi, Saitama 354, Japan

‡ National Institute of Health, Musashi-Murayama, Tokyo 190-12, Japan

Some wild African green monkeys are known to be naturally infected with a retrovirus related to human immunodeficiency virus (HIV) without having any apparent symptoms of an AIDS-like disease^{1,2}. This simian immunodeficiency virus, designated SIV_{AGM}, may be helpful in clarifying the evolution and pathogenicity of HIV. Some virus strains that were previously reported to be isolated from African green monkeys³⁻⁵ were shown to be laboratory contaminations of SIV_{MAC} (SIV from a rhesus macaque)⁶⁻⁸. Here we report the complete DNA sequence of authentic SIV_{AGM}, which was isolated from a naturally infected African green monkey of Kenyan origin. Comparison of the genome of SIV_{AGM} with those of known HIV/SIVs indicates that the virus is a new simian lentivirus that is approximately equally distantly related to HIV-1 and HIV-2 in contrast to SIV_{MAC}, which is much closer to HIV-2 than to HIV-1 (refs 5, 9).

The complete nucleotide sequence was obtained from a λ clone containing the full-length viral DNA derived from the circular replicative intermediate DNA of SIV_{AGM} (Fig. 1). This strain, SIV_{AGM}[TYO-1], was previously isolated in our laboratory from a naturally infected African green monkey imported from Kenya and was originally referred to as SIV[AGM-1] (ref. 2).

The genome of SIV_{AGM} is 9,170 nucleotides (nt) long (in its RNA form), which is shorter than the genomes of HIV/SIVs reported previously^{5,9-15}. The genetic organization of SIV_{AGM} (see Fig. 2), 5'LTR-gag-pol-central region-env-F(3'orf)-3'LTR, is typical of HIV/SIV group lentiviruses. The central region of SIV_{AGM} includes four open reading frames, Q (sor), tat, art (trs) and X, which can be identified by alignment with known HIV/SIVs. There is no detectable homologue of the R gene that is found in HIV-1, HIV-2 and SIV_{MAC}, however.

The SIV_{AGM} LTR (long terminal repeat) is 725 base pairs (bp) long. The boundaries of the U3-R and R-U5 segments within the LTR were determined by comparison between the sequences of an SIV_{AGM} complementary DNA clone and other HIV/SIVs. The sizes of the LTR segments U3, R and U5 are 507, 117 and 101 bp, respectively. The consensus promoter sequence (C/TATAA/TA) in U3 and the polyadenylation signal (AATAAA) in the LTR R sequence were identified, as were a conserved sequence of the enhancer^{16,17} and binding regions for the transcription factor Sp1^{16,18}. The enhancer sequences of SIV_{AGM} with two complete copies of GGGACTTTCCA (8,950-8,960 and 8,966-8,976) and three Sp1-like binding sequences (8,978-9,006) are more similar to those of HIV-1 than to those of HIV-2 or SIV_{MAC}.

The gag open reading frame of SIV_{AGM} starts at nt 432 and encodes 519 amino acids (total relative molecular mass (M_r), 58, (58.1 K)). The final gag gene products, p17, p24 and p15, of SIV_{AGM} were detected on Western blots². Alignment of the amino-acid sequences of SIV_{AGM} and other HIVs suggests that the p17 homologue contains 141 amino acids (M_r , 16K) and is probably generated by cleavage between Phe and Pro at nt852-857, as proline is the first amino acid at the N-terminus of both HIV-1 p24 (ref. 19) and HIV-2 p26 (ref. 15). The overall homologies of the gag precursors of SIV_{AGM} to those of HIV-1_{BRU} (ref. 10) and HIV-2_{ROD} (ref. 15) are 52.7% and 58.7%, respectively (Table 1).

The pol region in the SIV_{AGM} genome, nt 1,634-4,816, could encode 1,061 amino acids, and overlaps the gag gene by 355 bases. This open reading frame (total M_r , 120.6K) probably encodes three enzymes, the protease, reverse transcriptase and endonuclease. The junction between reverse transcriptase and the endonuclease of SIV_{AGM} is presumed to lie between the Val-Leu and Phe-Leu codons at nt 3,938-3,949, as this junction is conserved in HIV-1 (ref. 20), HIV-2 (ref. 15) and SIV_{MAC} (refs 5, 9). The overall amino-acid homology of the pol region of SIV_{AGM} with those of HIV-1_{BRU} and HIV-2_{ROD} is 56.5% in both cases. These relatively high homologies of the gag and pol regions of SIV_{AGM} with those of other HIV/SIVs probably accounts for the antigenic cross-reactivity between SIV_{AGM} and other HIV/SIVs².

Four open reading frames were identified in the central region

Table 1 Amino acid homology (%) between virus isolates

SIV _{AGM} compared with:	<i>gag</i>	<i>pol</i>	<i>env</i>	(EGP)*	(TMP)*	Q	R	X	<i>tat</i>	<i>art</i>	F
SIV _{MAC}	57.1	55.8	46.4	47.2	45.5	36.2	—	39.4	38.8	48.8	39.6
HIV-2 _{ROD}	58.7	56.5	44.1	43.4	47.2	39.0	—	36.7	44.8	47.7	39.4
HIV-1 _{BRU}	52.7	56.5	36.4	32.9	42.7	32.2	—	—	33.3	45.1	39.4
SIV _{MAC} compared with:											
HIV-2 _{ROD}	86.2	81.0	71.5	71.9	72.8	73.0	70.7	86.5	59.8	62.6	60.1
HIV-1 _{BRU}	53.1	56.7	36.0	32.5	42.2	30.5	60.7	—	47.3	34.3	37.4

The number indicates the % homology within the aligned region using Wilbur and Lipman programs²⁹ (parameters; KTPL 2-3, DEL 3-6).

* EGP and TMP are cleavage products of a protein precursor encoded by the *env* open reading frame.

[illegible]

Fig. 1 (Left and opposite) Complete nucleotide sequence of the SIV_{AGM} viral DNA genome. The sequence is numbered from the beginning of the LTR R region (cap site) to the end of the second LTR R region (polyadenylation site). In the LTR, the short inverted repeat (5'ACTG-CAGT3') is indicated by a horizontal arrow, and the promoter (CATAAA) and the polyadenylation signal (AATAAA) are boxed. The primer binding site (PBS) which is complementary to the tRNA_{Lys}, and two copies of polypurine tracts (PPT) are underlined. SD and SA indicate the splice donor and acceptor sites in the *tat* and *art*/*tr*s genes. **Methods.** The full length clone of SIV_{AGM}, λSAH12, was derived from extra-chromosomal closed-circular DNA of SIV_{AGM} (M.F. *et al.*, manuscript submitted). The insert was sequenced by the method by Sanger *et al.*²⁸

of SIV_{AGM} by its alignment with other HIV/SIVs, even though the homology with other HIV/SIVs is low (32.2–48.8%). The first of these, Q(*sor*) encodes a polypeptide of 235 amino acids, which is longer than those of other HIV/SIVs at its carboxy terminus. The second short open reading frame in the central region of SIV_{AGM} encodes a polypeptide of 119 amino acids. The length and amino-acid alignment of this peptide suggest that this open reading frame should be identified as X. The other two open reading frames, which overlap each other and are organized as split genes, are *tat* and *art(trs)*, which encode peptides of 100 and 84 amino acids, respectively. The peptides of SIV_{AGM} *tat* and *art(trs)* both share low amino-acid homology with those of other HIV/SIVs, but the Cys-rich region, the following Arg/Lys-rich region of the first coding exon of *tat* and the Arg-rich region of the second coding exon of *art(trs)* are relatively well conserved in these HIV/SIVs, suggesting conservation of essential structures and functions.

The *env* open reading frame of SIV_{AGM} encodes a precursor that is cleaved into an external glycoprotein (EGP) and a transmembrane protein (TMP) (Fig. 3). The calculated $M_{r,s}$ of the non-glycosylated products of EGP and TMP are 58.5K and 37.5K, respectively; the latter may be reduced in size to 26.5K by the in-frame stop codon in TMP at nt 8,077–8,079. This stop codon was previously found in HIV-2 (ROD35 isolate¹⁵) and SIV_{MAC} (refs 5, 9) at the splice acceptor site of the *tat* and *art* (*trs*) genes, but the stop codon of SIV_{AGM} is 96 nt downstream from this splicing acceptor site. The homology of the amino-acid sequence of SIV_{AGM} with those of other HIV/SIVs is relatively low in the *env* region (36.4–46.4%), and SIV_{AGM} shows slightly higher homology with HIV-2/SIV_{MAC} than with HIV-1. In addition, of the 30 cysteines present in the SIV_{AGM} *env* protein, 23, 24 and 18 are found at the same positions in SIV_{MAC}, HIV-2_{ROD} and HIV-1_{BRU}, respectively. Furthermore, the 24 N-glycosylation sites in SIV_{AGM}, 14, 15 and 8 are identical to those in SIV_{MAC}, HIV-2_{ROD} and HIV-1_{BRU}, respectively. These may account for the greater antigenic similarity to HIV-2/SIV_{MAC} than to HIV-1. The degree of sequence similarity between all of the HIV/SIV TMP sequences is high in the N-terminal but low in C-terminal regions (except in the region of the putative IL-2 homology at the C-terminus)⁴, which are roughly divided by the placement of the in-frame stop codon. The functional sequence which is related to membrane fusion activity in the N-terminus of TMP in HIV-1 (ref. 21) is conserved in SIV_{AGM}. The TMP N-terminal (off) amino-acid sequence Phe-Leu-Gly-Phe-Leu, which is identical to the F1 glycoprotein of respiratory syncytial virus (ReSV) is also conserved in SIV_{AGM} as well as in other HIV/SIVs²².

The last open reading frame, F (3'*orf*) of SIV_{AGM}, encodes a peptide of 229 amino acids. The conserved region of F, which includes the phosphorylated GTP-binding protein-like sequence²³, is found in SIV_{AGM}.

Comparisons of all the open reading frames of SIV_{AGM} with those of HIV-1, HIV-2 and SIV_{MAC} are summarized in Table 1. Based on these comparisons, our SIV_{AGM} is considered to be approximately equally distant from all other HIV/SIVs, although it is slightly closer to HIV-2/SIV_{MAC} than to HIV-1



Fig. 3 Comparison of the *env* proteins of SIV_{AGM}, SIV_{MAC}, HIV-2_{ROD} and HIV-1_{BRU}. Alignments were made using Wilbur and Lipman programs²⁹ with reference to published data (ref. 9). The sequences identical to SIV_{AGM} are indicated by asterisks, and gaps are shown as dashes. Cysteines are boxed, and potential *N*-glycosylation sites are underlined. The potential cleavage sites of HIV-1_{BRU} are shown by arrows. The termination codons in SIV_{AGM} and SIV_{MAC} TMP are indicated by filled circles.

The current arguments on the origin and transmission of HIV/SIVs remind us of an earlier controversy on HTLV-I, the aetiological agent of adult T-cell leukaemia, and its related simian retroviruses (STLV-Is). STLV-Is are genetically distinguishable from HTLV-I and are species-specific in non-human

Another point of interest is the difference in pathogenicity of HIV/SIV group viruses. Some regions of the present SIV_{AGM}

isolates are closely related to HIV_s, and *in vitro* infection with SIV_{AGM} induces remarked cytopathic effects in human T lymphocytes². No SIV_{AGM}-infected African green monkeys have yet been found to show clinical signs of illness, however, and SIV_{AGM} can infect and grow well in peripheral blood lymphocytes from African green monkeys but induces no cytopathic effects. We think the lack of pathogenicity of SIV_{AGM} in African green monkeys may be related to the genetic characteristics of SIV_{AGM}, such as the lack of an R open reading frame in the central region and the in-frame stop codon in TMP. Comparative analyses of the biological functions of the SIV_{AGM} and pathogenic HIV/SIVs will be necessary to elucidate the reason for the differences in pathogenicity.

Received 10 March; accepted 18 April 1988.

1. Kanki, P. J., Alroy, J. & Essex, M. *Science* **230**, 951-954 (1985).
2. Ohta, Y. *et al. Int. J. Cancer* **41**, 115-122 (1988).
3. Hirsch, V. *et al. Proc. natn. Acad. Sci. U.S.A.* **83**, 9754-9758 (1986).
4. Hirsch, V., Riedel, N. & Mullins, J. I. *Cell* **49**, 307-319 (1987).
5. Franchini, G. *et al. Nature* **328**, 539-543 (1987).
6. Newmark, P. *Nature* **326**, 548 (1987).
7. Desrosiers, R. C., Daniel, M. D., Letvin, N. L. & Hunt, R. D. *Nature* **327**, 107 (1987).
8. Kestler III, H. W. *et al. Nature* **331**, 619-621 (1988).
9. Chakrabarti, L. *et al. Nature* **328**, 543-547 (1987).
10. Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. & Alizon, M. *Cell* **40**, 9-17 (1985).
11. Ratner, L. *et al. Nature* **313**, 277-284 (1985).
12. Muesing, M. A. *et al. Nature* **313**, 450-458 (1985).
13. Sanchez-Pescador, R. *et al. Science* **227**, 484-492 (1985).
14. Alizon, M., Wain-Hobson, S., Montagnier, L. & Sonigo, P. *Cell* **46**, 63-74 (1986).
15. Guyader, M. *et al. Nature* **326**, 662-669 (1987).
16. Kornfeld, K., Riedel, N., Vigilanti, G. A., Hirsch, V. & Mullins, J. I. *Nature* **326**, 610-613 (1987).
17. Nabel, G. & Baltimore, D. *Nature* **326**, 711-713 (1987).
18. Jones, K. A., Kadonaga, J. T., Luciw, P. A. & Tjian, R. *Science* **232**, 755-759 (1986).
19. Casey, J. M. *et al. J. Virol.* **55**, 417-423 (1985).
20. Lightfoote, M. M. *et al. J. Virol.* **60**, 771-775 (1986).
21. Kowalski, M. *et al. Science* **237**, 1351-1355 (1987).
22. Gallaher, W. R. *Cell* **50**, 327-328 (1987).
23. Guy, B. *et al. Nature* **330**, 266-269 (1987).
24. Wong-Staal, F. *et al. Science* **229**, 759-762 (1985).
25. Clavel, F. *et al. Nature* **324**, 691-695 (1986).
26. Komuro, A. *et al. Virology* **138**, 373-378 (1984).
27. Hayami, M. *Cancer Rev.* **1**, 35-63 (1986).
28. Sanger, F., Nicklen, S. & Coulson, A. R. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463-5467 (1977).
29. Wilbur, W. J. & Lipman, D. J. *Proc. natn. Acad. Sci. U.S.A.* **80**, 726-730 (1983).

Demonstration of somatic mutation and colonic crypt clonality by X-linked enzyme histochemistry

D. F. R. Griffiths, S. J. Davies, D. Williams,
G. T. Williams & E. D. Williams

Department of Pathology, University of Wales College of Medicine,
Heath Park, Cardiff CF4 4XN, UK

Cellular mosaicism resulting from X-chromosome inactivation in heterozygous females¹ can be shown histochemically²⁻⁴; using this approach we have demonstrated age-related gene reactivation⁵ and tumour clonality⁶. We now show in female mice heterozygous for reduced expression of glucose-6-phosphate dehydrogenase (G6PD) activity that colonic epithelial cells express either normal or low enzyme activity, and form patches composed of multiple crypts of uniform phenotype. We also show that a low-enzyme colonic epithelial cell phenotype can be induced in normal mice by carcinogen treatment, these cells again occur in patches, but are restricted to scattered single crypts, the frequency of which is related to treatment. A small proportion of colonic tumours in carcinogen treated normal mice are also of low-enzyme phenotype. We conclude that we have visualized the effects of a sporadic carcinogen induced somatic mutation in the *G6PD* gene of crypt stem cells and that a single stem cell maintains each colonic crypt. This inducible defective activity of a ubiquitous 'housekeeping' enzyme provides a somatic clonal marker system of wide potential application.

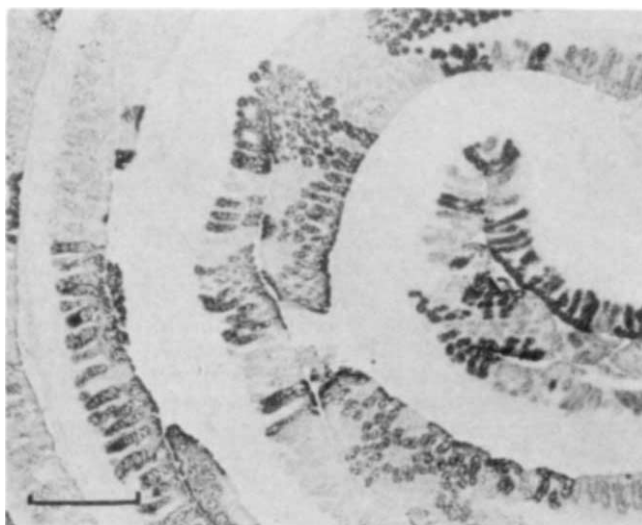


Fig. 1. Mosaic pattern of histochemical localization of G6PD activity in large bowel from an untreated female mouse heterozygous for the *GPDx* gene (C3H×GPDx). The picture is representative of that seen in ten 12-week-old animals. They were killed by CO₂ inhalation, the bowel removed, opened, coiled to form a 'Swiss roll' and frozen in isopentane cooled to -70 °C with dry ice. G6PD activity was demonstrated in 6 µm sections by incubation in media containing 5 mM nitroblue tetrazolium, 2.09 mM NADP, 5 mM G6P, 4 mM MgCl₂, 5 mM NaN₃, 0.32 mM phenazine methosulphate and 32% polyvinyl alcohol GO4/140 in 0.05 M Tris-HCl buffer at pH 7.5 for 30 minutes at 37 °C. After post-fixation in 4% formal saline and washing in water, sections were mounted in glycerine jelly. The specificity of the reaction was confirmed by the use of control sections incubated in media devoid either of G6P alone or of both G6P and NADP. G6PD histochemistry, frozen section. Scale bar, 0.5 mm.

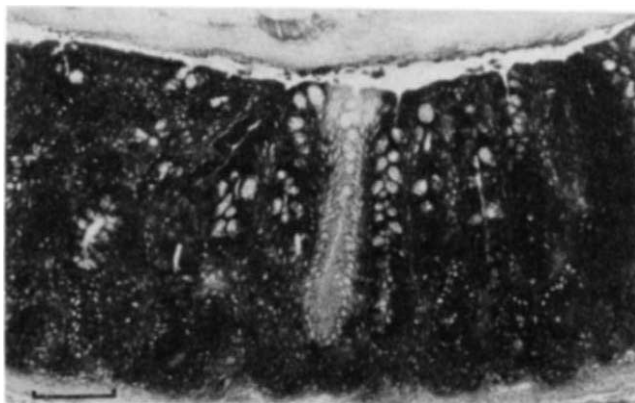


Fig. 2 A single crypt showing low G6PD activity in a background of crypts expressing normal levels of enzyme activity in a female mouse homozygous for the wild-type *G6PD* gene but treated for 21 weeks with weekly subcutaneous injections (20 mg per kg) of the carcinogen dimethyl hydrazine. Crypts are of normal morphology (confirmed by H and E staining of adjacent sections). The appearances are typical of those seen in treated TO strain and C3H strain mice. The animals were killed two weeks after the last dose of carcinogen; G6PD has been demonstrated by the method described in Fig. 1. G6PD histochemistry, frozen section. Scale bar 0.05 mm.

G6PD (EC. 1.1.1.49) is an X-linked enzyme in the mouse as it is in man⁷. We have used the histochemistry of G6PD to study clonality in the large intestinal mucosa of 40 female mice, 10 normal C3H, 10 normal TO, 10 G6PD deficient (GPDx)⁸ and 10 C3H×GPDx, using whole colon sections cut at 10 levels. In normal mice all colonic epithelial cells uniformly expressed

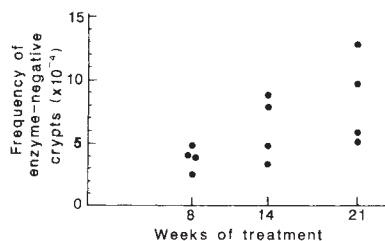


Fig. 3 Frequency of single crypts with low G6PD activity related to DMH treatment. Female TO mice (12 weeks old) were treated with 8, 14 and 21 weekly subcutaneous injections of the carcinogen DMH (20 mg per kg) and killed two weeks after the last dose. Four treated, and four age- and sex-matched control animals were studied for each time point. 'Swiss rolls' of the entire colon for each animal were sectioned at 12 levels at least 50 μ m apart and stained for G6PD activity (method as in Fig. 1). The number of crypts with low enzyme activity in 12 sections from each colon was counted by an observer unaware of the treatment schedule. The total numbers of enzyme negative crypts in each animal were: after 8 weeks' treatment 18, 21, 20 and 21; after 14 weeks 48, 18, 32 and 45; and after 21 weeks 34, 33, 51 and 62. The total number of large bowel crypts in the counted sections from each animal was estimated by counting crypts in 20 fields of 0.458 mm² selected by a battlement protocol and relating this to the total area of section measured on an IBAS II image analyser. Each point shows the frequency ($\times 10^{-4}$) of enzyme-negative crypts. There is a significant correlation ($P < 0.02$) between the frequency of enzyme-negative crypts and the number of weekly injections (Spearman's rank correlation). No low activity crypts were found in an equal number of untreated animals examined at 22, 28 and 35 weeks of age.

high levels of G6PD activity, whereas in G6PD mice they were of uniformly low activity. All heterozygous mice showed large patches of high and low activity epithelial cells (Fig. 1). The lack of mixed activity crypts suggests that during development all epithelial cells in a colonic crypt are derived from a single cell in biparental animals, as has been found in tetraparental animals⁹. Both columnar and goblet cells within a crypt showed similar levels of cytoplasmic enzyme activity. Enterro-endocrine cells are smaller and difficult to identify in frozen section, we cannot comment on their lineage.

Administration of the colon-specific carcinogen dimethyl hydrazine (DMH) to C3H mice consistently led to the appearance of epithelial cells of low G6PD activity in patches restricted to individual crypts of normal morphology (Fig. 2). In these 'negative crypts' the whole crypt was made up of epithelial cells with the low-activity phenotype. These cells ceased abruptly on the surface epithelium, approximately at the midpoint between the affected and the neighbouring unaffected crypt. This change was similar to the low enzyme activity seen in the heterozygous

G6PD-deficient mice in that it was crypt restricted and no crypts showing mixed low and normal enzyme activity were seen. It differed in that solitary enzyme-negative crypts were widely scattered apparently at random throughout the whole of the large intestine rather than occurring in multiple crypt patches. The effect was quantified by counting the number and frequency of enzyme-negative crypts in whole colon sections of 12 TO mice treated with up to 21 weekly doses of DMH (Fig. 3). The frequency of affected crypts rose with time; a mean of 3.6×10^{-4} crypts was found after 8, 6.2×10^{-4} after 14, and 8.4×10^{-4} after 21 weekly doses.

This carcinogen-induced low-enzyme phenotype occurs in clearly defined small patches conforming to the growth pattern of the intestine, and is therefore heritable. It is induced in a dose-dependent way by a carcinogen known to be a mutagen¹⁰. It affects a housekeeping gene in homozygous animals and is not associated with any morphological abnormality or lack of differentiation. We therefore believe that this change is due to a somatic mutation. As individual crypts only were affected and the phenotype change was uniform within the affected crypts, we suggest that the carcinogen-induced mutation occurred in the *G6PD* gene of a single stem cell giving rise to all the epithelial cells in the affected crypt. However cell kinetic studies estimate that intestinal crypts arise from the continued proliferation of a number of stem cells (up to 20 for the small intestinal crypt, and 1–3 for the large intestinal crypt^{11,12}). We studied 12 whole colon sections from each of 28 carcinogen-treated mice containing in all ~500,000 crypt profiles, and failed to find a single crypt affected in the asymmetrical patchy way expected if there were multiple crypt stem cells. We therefore conclude that the epithelial cells of large intestinal crypts are not only derived from a single cell during embryonic development, but are also each maintained by a single stem cell.

The phenotype of tumours was investigated in five C3H and 11 TO mice treated with DMH for up to 28 weeks. A total of 141 intramucosal tumours¹³ were found; 134 showed the expected uniform enzyme positivity. Two tumours, however, were entirely negative for G6PD (Fig. 4), whereas in five there were negative patches. All animals showed scattered negative crypts in the background mucosa. We again interpret these results as due to carcinogen-induced mutation in the *G6PD* gene; mutation in the cell of origin giving rise to an entirely negative tumour, and mutation in the cells of early tumours giving rise to negative patches. The occurrence of such small numbers of entirely negative tumours is compatible with a single-cell or single-crypt origin of large intestinal tumours in these normal animals, and is in agreement with the finding in tetraparental animals¹⁴.

Histochemical methodology provides a hitherto unused approach to study carcinogen induced mutation. We have used an X-linked enzyme, and therefore one in which the mutation

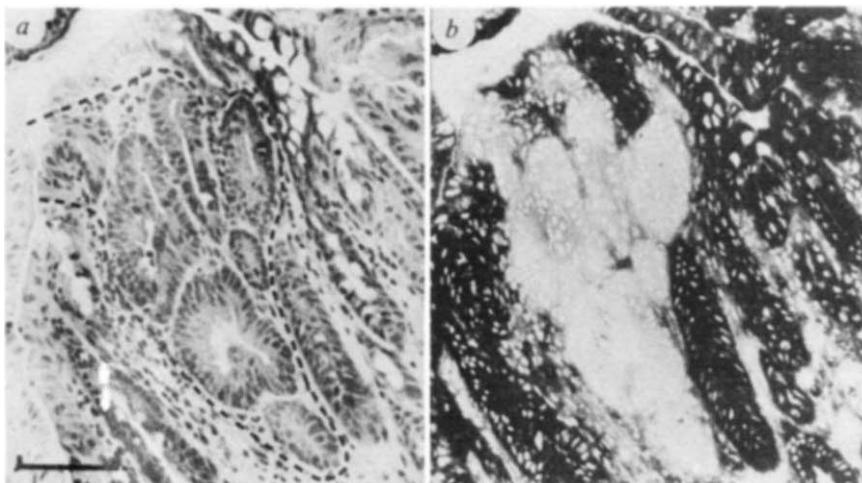


Fig. 4 *a*, A small intramucosal tumour (outline indicated) induced in a female mouse homozygous for wild-type G6PD activity by 27 weekly injections of DMH (15 mg per kg). H & E, frozen section. *b*, Serial section stained for G6PD shows low enzyme activity within neoplastic tubules. Surrounding non-neoplastic crypts show normal enzyme activity. G6PD, frozen section. Scale bar, 0.1 mm.

of the gene on one allele will not normally be compensated at a cellular level by the activity of the gene on the other allele. Other X-linked enzymes could provide alternative markers and show whether there is any specificity in the pattern of enzyme loss induced by different carcinogens. The technique is of particular value in studying the size and distribution of clonal units. Calculation of the carcinogen-induced mutation rate in stem cells requires a more precise knowledge of their kinetics than is currently available for the intestine. However the crypt conversion rate per dose of carcinogen varied in different animals between 0.25×10^{-4} and 0.64×10^{-4} . We saw no example of spontaneous crypt conversion in 21 untreated animals between 14 and 35 weeks of age; the sections studied contained over 400,000 crypts, suggesting that the spontaneous mutation rate is low.

The ability to visualize a sporadic heritable carcinogen-induced mutation leading to enzyme loss in somatic cells in adult animals provides a valuable inducible clonal marker system which has wide potential applications, not only in developmental biology, but also in studies of both experimental and human carcinogenesis.

We thank the Cancer Research Campaign and the Welsh Scheme for Development of Health and Social Research for financial support.

Received 9 March; accepted 21 April 1988.

1. Lyon, M. F. *Nature* **191**, 372-373 (1961).
2. Wareham, K. A., Howell, S., Williams, D. & Williams, E. D. *Histochem. J.* **15**, 363-371 (1983).
3. Thomas, G. A., Williams, D. & Williams, E. D. *J. Pathol.* (in the press).
4. Lyon, M. F. *et al. J. Embryol. exp. Morph.* **97**, 75-85 (1986).
5. Wareham, K. A., Lyon, M. F., Glenister, P. H. & Williams, E. D. *Nature* **327**, 725-727 (1987).
6. Howell, S., Wareham, K. A. & Williams, E. D. *Am. J. Pathol.* **121**, 426-432 (1985).
7. Goss, S. J. *J. cell. Sci.* **72**, 241-257 (1984).
8. Pretsch, W., Charles, D. J. & Merkle, S. *Biochem. Genet.* **26**, 89-103 (1988).
9. Ponder, B. A. J. *et al. Nature* **313**, 689-691 (1985).
10. LaMont, J. T. & O'Gorman, T. A. *Gastroenterology* **75**, 1157-1169 (1978).
11. Bykorez, A. I. & Ivashchenko, Y. D. *Int. Rev. Cytol.* **90**, 309-373 (1984).
12. Potten, C. S. & Hendry, J. H. in *Stem Cells* (ed. Potten, C. S.) 155-199 (Churchill Livingstone, Edinburgh, 1983).
13. Chang, W. W. L. *Virchow's Arch B* **41**, 17-37 (1982).
14. Ponder, B. A. J. & Wilkinson, M. M. *J.N.C.I.* **77**, 967-976 (1986).

A clonal marker induced by mutation in mouse intestinal epithelium

D. J. Winton, M. A. Blount & B. A. J. Ponder

Section of Human Cancer Genetics, Institute of Cancer Research, Haddow Laboratories, Clifton Avenue, Sutton, Surrey SM2 5PX, UK

A cellular marker for individual somatic cells and their clonal descendants would be a valuable tool for the investigation of cell lineages and clonal organization in developing and in renewing tissues. Such markers have been developed in *Drosophila*¹, but (apart from mutant melanocytes in retinal pigmented epithelium^{2,3}) not so far in mammalian tissues. We report here the development of a mutation-induced marker in mice heterozygous at the *Dlb-1* locus which determines the expression of binding sites for the lectin *Dolichos biflorus* agglutinin (DBA) in intestinal epithelium⁴. We show that this marker can be used to study the clonal organization of adult intestinal epithelium, and to mark descendent clones arising during development. The method can in principle be extended to any other suitable markers which can be obtained in a heterozygous state, including markers generated in transgenic animals.

Previous analysis of the cellular mosaicism in embryo aggregation chimaeras⁵⁻⁷ have shown that their use in cell lineage studies is limited by the marking of the mosaic cell population from the earliest stages of embryonic development. As a result, in adult epithelia, the patch size of the mosaic is in general far larger than the affected territory of a single renewing stem cell and its descendants, making, for example, analysis of the clonal



Fig. 1 Clonal patterns in intestinal epithelium. *a*, Jejunal villus from a C57BL/6J x SWR F₁ mouse which had received ENU (50 mg per kg) two weeks before killing. Dissected from a whole mount stained with DBA peroxidase⁷. A ribbon of unstained (DBA-negative) epithelial cells is seen (arrowed) similar in appearance to those observed in embryo aggregation chimaeras (Fig. 1c; ref. 7). Each ribbon presumably arises from mutation within the stem cell compartment as at least 70 mutated cells would be needed to make an unbroken ribbon of single cell width stretching from the base to the tip of a jejunal villus¹². This would require at least six subsequent divisions of a mutated cell arising in the transit population ($2^6 = 64$), which is more than the four transit divisions indicated by models of crypt cell kinetics¹³, based on 16 stem cells per crypt. The irregular dark lines are folds in the villus epithelium. Bar = 100 μ m. *b*, Cross-section of a villus bearing a DBA-negative ribbon of five epithelial cells (arrowed) similar to that shown in *a*. C57BL/6J x SWR F₁ mouse treated with ENU as in *a*. The mutated clone was located in a whole mount preparation stained with DBA peroxidase and excised with a block of surrounding tissue. After fixation in methacarn and embedding in paraffin, 4- μ m sections were cut, restained with DBA peroxidase and counterstained with haemalum. Bar = 100 μ m. *c*, Jejunal villus of a C57BL/6J x SWR chimaera¹⁰ dissected from a small intestinal whole mount and stained with DBA peroxidase as in *a*. An unstained ribbon of SWR epithelium is seen ascending the villus (arrow). Bar = 100 μ m.

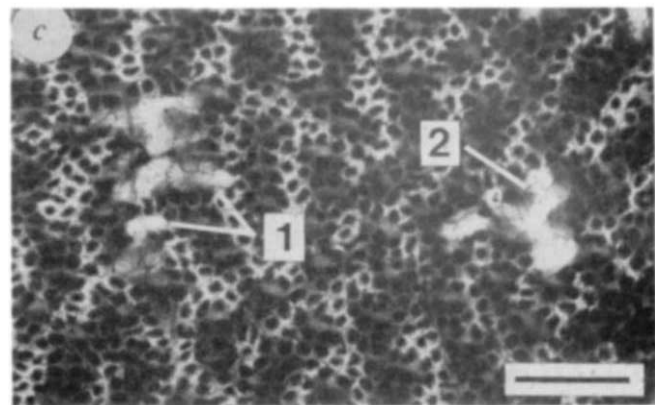
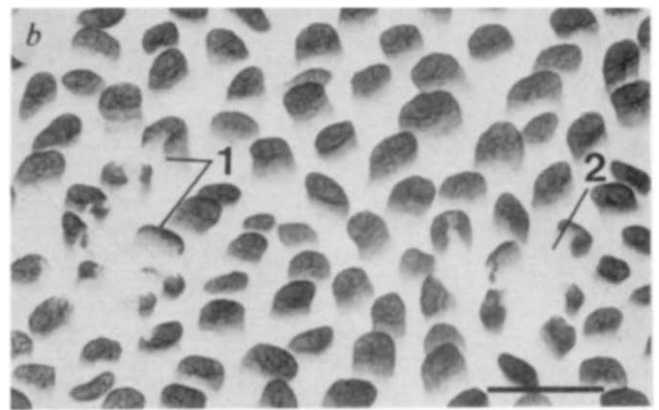
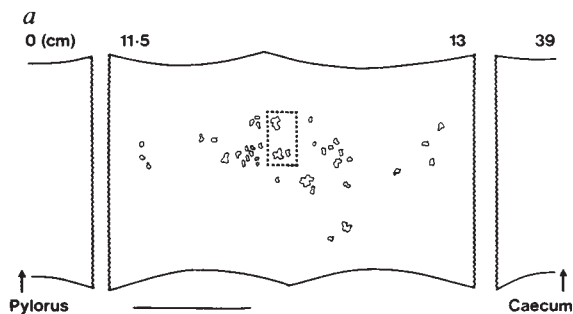


Fig. 2 A descendent clone of mutated crypts induced by *in utero* exposure of a C57BL/6J×SWR F_1 embryo to ENU. *a*, Diagrammatic representation of a cluster of DBA-negative crypts observed in a C57BL/6J×SWR F_1 mouse killed at 8 weeks of age and exposed to ENU while *in utero* at day 8 of gestation. Bar = 0.5 cm. Each of the outlined areas represents one or more DBA-negative crypts. The pinned out length of the small intestine was 39 cm with the descendent clone involving 1.5 cm (3.8% of total length) of jejunum lying between 11.5 and 13-cm distal from the pylorus, and containing an approximate total of 300 negative crypts. Similar clones are seen in chimaeric intestine¹¹. *b*, *c*, Higher power micrographs (Bar = 0.5 mm) of two DBA-negative patches (arrowed 1 and 2) from enclosure shown in *a*, box rotated through 90°, and viewed from *b* the luminal (villus) side and (*c*) the abluminal side: the external muscle coats of the intestine have been stripped away to reveal the crypt bases. Patches 1 and 2 involve 10 and 8 villi and arise from approximately 50 and 30 negative-staining crypts respectively.

Methods. C57BL/6J female mice were mated with SWR males and checked the following day for copulation plugs. Mice with plugs were housed separately and injected at day 8 of gestation (day of plug = day 1) with ENU (50 mg per kg) as described in Fig. 3. Mice were kept under standard laboratory conditions (legend to Fig. 3) until 8 weeks of age when they were killed and descendent clones of DBA-negative epithelium were identified in whole mount preparations as described in Fig. 1. To view the pattern of DBA-negative staining epithelium from the abluminal (crypt base) side, a length of intestine containing the negative staining patch was turned over and refixed in a methanol/glacial acetic acid (85:15) mixture before storage in 70% ethanol. This facilitated stripping of the muscle layers from the remainder of the epithelium. After rehydration through a graded ethanol series, crypt bases were stained with DBA-peroxidase conjugate as described in Fig. 1.

organization within the intestinal crypt impossible⁵.

To mark smaller clones of cells at a chosen time, we have used C57BL/6J×SWR F_1 mice, which are heterozygous *Dlb-1^b/Dlb-1^a*, instead of chimaeras. C57BL/6J mice (*Dlb-1^b*) express DBA binding sites on intestinal epithelium but not on vascular endothelium; SWR mice (*Dlb-1^a*) show the reverse pattern^{4,8}. DBA binding is co-dominant: in F_1 mice both tissues express the binding site. Backcross experiments show that the different patterns of expression are due to alleles at a single locus, mapped to chromosome 11 (refs 4, 8). The F_1 mice have only one allele specifying the expression of the DBA binding site on intestinal epithelium. Loss of this allele by mutation or somatic recombination in a progenitor or stem cell should lead to a clone of cells which do not express the DBA binding site, and are not stained using a DBA-peroxidase conjugate. From our studies in chimaeras, we predicted that a clone arising from such a mutation in a crypt stem cell should be recognizable in small intestinal epithelium as a ribbon of DBA-negative cells extending up a villus. Mutations occurring before the time of crypt formation (which is immediately after birth) should give rise to clusters or 'descendent clones'⁹ of wholly DBA-negative crypts in the adult animal. Each of these patterns is indeed found in untreated F_1 mice and, at a higher frequency, in mice that have been treated with the mutagen ethylnitrosourea (ENU) (Figs 1 and 2).

The evidence that these patterns are the result of somatic mutation is fourfold: (1) resemblance to the clonal patterns seen in chimaeric tissue (Figs 1 and 2); (2) absence in C57BL/6J homozygous *Dlb-1^b/Dlb-1^b* control mice (no DBA-negative ribbons were observed in six untreated C57BL/6J mice, nor in

11 C57BL/6J mice treated with ENU 14 days previously (10^4 jejunal villi scored in each mouse)); (3) increased frequency after exposure to the mutagen ENU in a dose-dependent way (Fig. 3a); (4) frequency consistent with somatic mutation.

Two weeks after a single dose of ENU (100 mg per kg), 55 ribbons were observed per 10^4 villi (Fig. 3a). Subtracting the background level found in controls (4 ribbons per 10^4 villi) and assuming that 10 crypts contribute to each jejunal villus¹² and that there are 4–16 cycling stem cells per crypt, mutation in any one of which might give rise to ribbons¹³, gives an event rate of 51 per 10^5 crypts, or 3.2–12.8 per 10^5 cells. The spontaneous mutation rate can be estimated from the accumulation of villus ribbons with age in untreated F_1 mice (Fig. 3b). The number of ribbons present at any time will reflect both the accumulation of new mutations up to that time, and their loss as stem cells are replaced. That the increase in number of ribbons is linear up to 26 weeks of age suggests a constant mutation rate and minimal stem cell loss during this period: that is, the stem cells which give rise to the ribbons have a 'tenure' in the crypt of at least six months, approximately one quarter of the lifespan of the mouse. The slope of the line in Fig. 3b indicates that, in the first six months of life, the incidence of presumed mutations is 1 per 10^4 villi per 4 weeks, or 1 per 28×10^5 crypts per day. Assuming 8–13 mitoses per 24 h in the crypt stem cell compartment¹³ gives an estimated spontaneous mutation rate of 3.0–4.5 per 10^8 mitoses. The types and frequency of genomic alterations which lead to DBA-negative clones cannot be determined directly, but there is evidence that ENU produces minimal gross chromosomal damage¹⁴ and acts primarily through the induction of point mutations¹⁵.

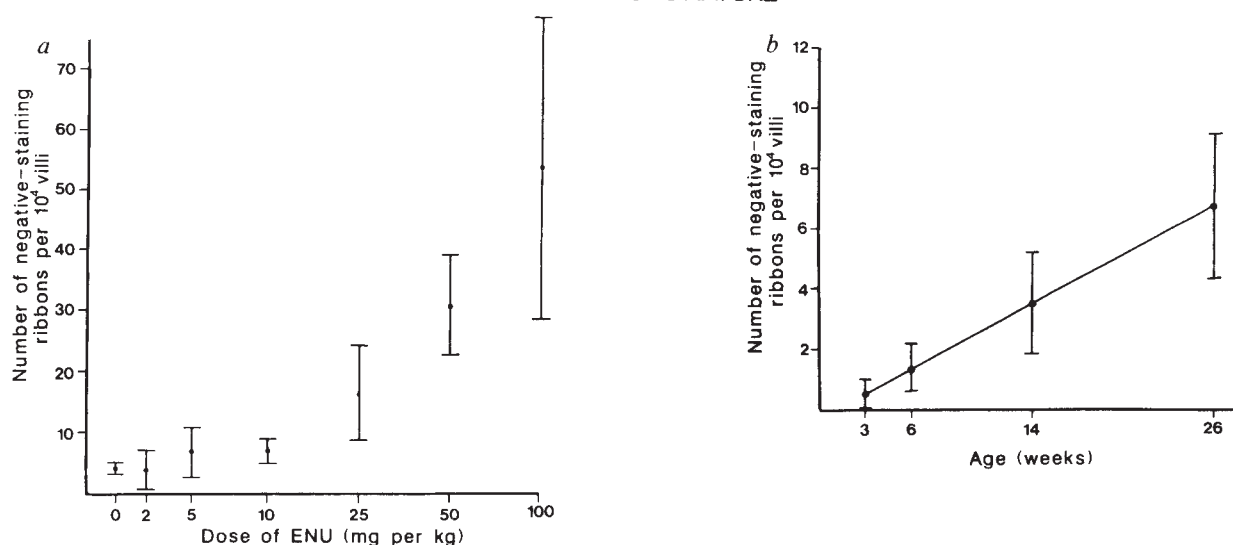


Fig. 3 Incidence of DBA-negative staining ribbons on the villi of C57BL/6J x SWR F₁ mice. *a*, Dose-response relationship with ENU. *b*, Accumulation with age in untreated F₁ mice.

Methods. Animals maintained under standard conditions (12-h night/day cycle, and water and SDS No 1 (modified) expanded diet (Special Diet Services, Essex, UK) supplied *ad libitum* were taken at random and killed by cervical dislocation. Whole mount preparations (see legend to Fig. 1) of the entire small intestine were stained with DBA peroxidase. All whole mounts were randomly numbered and scored blind. Scoring of villi started at a point 25% of the length of intestine from the pylorus. Villi were examined using a Kyowa Stereo 200M microscope at a total magnification of $\times 45$. Fifty fields delineated by the square field of an eyepiece graticule (Agar Aids, L4054), each containing ~ 200 villi, were scored for the presence of villus ribbons (Fig. 1a) not staining with the DBA-peroxidase conjugate. The fields were contiguous except where occasional damaged areas of the specimen made scoring impossible. Multiple negative ribbons on neighbouring villi arising from an apparent common centre were scored as one event. Duplicate counts were made of the number of villi in the first and last fields in the analysis, and the total number of villi scored (usually between 10,000 and 12,000) was estimated from the mean of these four observations. For each animal the final number of events was corrected to 10^4 villi using the formula: Total number of events $\times [10,000/(\text{Mean number of villi per field} \times 50)]$. *a*, For the dose-response curve, ENU (Isopac, N3385, Sigma) was dissolved in 0.9% sterile saline to form a stock solution (10 mg ml^{-1}) which was used within a week of reconstitution. Immediately before use, stock solutions were diluted in sodium phosphate buffer (pH 7.6, 0.12 M) to give a working solution (pH 6.8) of 3.3 mg ml^{-1} . In mice receiving 2, 5 and 10 mg per kg the working solution was further diluted with 0.9% sterile saline to give a minimum 0.1 ml for injection. Mice, 14–18 weeks old, were allocated to treatment groups without regard to sex. All animals were injected intraperitoneally (i.p.) within a few minutes with the same batch of ENU at $\sim 1500 \text{ h}$. Mean scores of villus ribbons were derived for each dose from 6 (0–10 mg per kg) or 4 (25–100 mg per kg) animals killed between 13 and 15 days after treatment. *b*, Data for the spontaneous accumulation of mutations was derived from 12 mice for each time point. The mice used for the 3, 6 and 14-week time periods were one cohort, and mice from the 26-week time point an earlier cohort, from the same breeding stock. All the experimental animals were killed and scored within the same 12-week period. Results are presented as mean $\pm$ s.d.

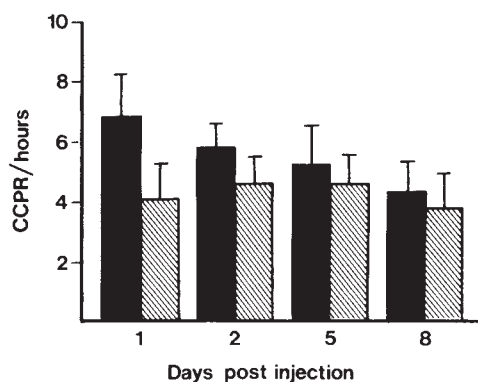


Fig. 4 Effect of ENU on jejunal crypt cell production rate (CCPR). Two groups of 40 C57BL/6J x SWR F₁ mice (12–16 weeks old) were injected i.p. with either ENU (50 mg per kg) (filled bars) or sodium phosphate buffer (hatched bars) as described in the legend to Fig. 3. Ten mice from each group were sampled at the time points indicated. Mice were killed at 15-min intervals between 30 and 165 min following the injection of vincristine sulphate (1 mg per kg)¹⁶. A sample of tissue was taken from a site at 30% of the length of the small intestine from the pylorus and fixed in Carnoy's fluid. The number of arrested metaphases in twenty micro-dissected, Feulgen-stained crypts from each animal was counted¹⁶ and the mean value plotted against time after injection. The slope of the metaphase arrest line (fitted by least square regression) was the crypt cell production rate. CCPRs were compared using a computer based *t*-test and are presented as means $\pm$ s.e.m. Comparison of the treated and control groups at each time point showed no significant difference between the CCPRs of the two groups. An additional comparison of pooled treated and control groups using the Generalised Linear Interactive Modelling (GLIM) system computer analysis package also failed to show a significant effect of ENU on the CCPRs.

The use of mutagenesis to generate marked clones carries the inevitable risk of disrupting tissue organization by cell death or impaired proliferation. A single dose of ENU (50 mg per kg) does not, however, cause a significant increase in the crypt cell production rate (Fig. 4), which suggests that it does not cause extensive cell death. ENU is known to be highly mutagenic in relation to its cytotoxicity^{14,17} and this observation suggests that the induction of a somatic clonal marker by ENU can be achieved without major tissue disruption.

Studies of cell lineages in development present a greater problem. At early stages of gestation, target populations are small and the embryo may be especially sensitive to the cytotoxic effects of mutagens¹⁸. At later stages, specific tissues may be sensitive to carcinogenic effects¹⁹. In the present study, 19 C57BL/6J x SWR F₁ offspring of four C57BL/6J dams injected with ENU (50 mg per kg) at day eight of gestation showed DBA-negative descendent clones (for example, Fig. 2) but no gross intestinal or other abnormalities, although the average

litter size was somewhat lower than for untreated mice (4.7 compared with 5.2). These preliminary observations suggest that the somatic mutation method may have practical application for studies of intestine during mid-gestation.

The DBA marker can also be used to investigate the incidence and accumulation of spontaneous and induced mutations in somatic tissues. In this respect, it extends the principle of the mammalian 'spot test', an *in vivo* mutation assay which depends upon the production of spots of altered coat colour in adult mice by mutation of melanocytes or hair follicle cells *in utero*^{2,3,20,21}, to a target tissue (the intestinal epithelium) where there is a much larger population of cells at risk. In principle, the method can be extended to any other tissue for which a suitable cellular marker can be obtained in the heterozygous state. The constitutive expression of markers such as β -galactosidase, present in single copy in mouse strains derived by gene transfer, will extend the range of possibilities. The practical application of the method will be enhanced if it is also possible to construct back mutation systems using inactivated marker genes which can be reactivated *in vivo* with sufficient frequency to provide a positively marked clone in a negative background.

This work was supported by grants to the Institute of Cancer Research from the Cancer Research Campaign and MRC. We

thank Dr R. Goodlad for assistance with the determination of crypt cell production rates.

Received 9 March; accepted 22 April 1988.

1. Lawrence, P. A. & Martinez-Arias, A. *Phil. Trans. R. Soc. B* **312**, 83-90 (1985).
2. Searle, A. G. & Stephenson, D. A. *Mutation Res.* **92**, 205-215 (1982).
3. Stephenson, D. A. & Searle, A. G. *Mutagenesis* **1**, 135-141 (1986).
4. Ponder, B. A. J., Festing, M. F. W. & Wilkinson, M. M. *J. Embryol. exp. Morph.* **87**, 229-239 (1985).
5. Ponder, B. A. J. *et al. Nature* **313**, 689-691 (1985).
6. Schmidt, G. H., Wilkinson, M. M. & Ponder, B. A. J. *Cell* **40**, 425-429 (1985).
7. Schmidt, G. H., Wilkinson, M. M. & Ponder, B. A. J. *J. Embryol. exp. Morph.* **91**, 197-208 (1986).
8. Uiterdyk, H. G. *et al. Genet. Res.* **47**, 125-129 (1986).
9. West, J. D. in *Development in Mammals* Vol. 3 (ed. Johnson, M. H.) 413-460 (Amsterdam: North-Holland/Elsevier, 1978).
10. Ponder, B. A. J., Wilkinson, M. M. & Wood, M. J. *Embryol. exp. Morph.* **76**, 83-93 (1983).
11. Schmidt, G. H., Garbutt, D. J., Wilkinson, M. M. & Ponder, B. A. J. *J. Embryol. exp. Morph.* **85**, 121-130 (1985).
12. Wright, N. A. & Irwin, M. *Cell Tissue Kinet.* **15**, 595-609 (1982).
13. Potten, C. S. & Loeffler, M. J. *theor. Biol.* **127**, 381-391 (1987).
14. Russell, L. B. & Montgomery, C. S. *Mutation Res.* **92**, 193-204 (1982).
15. Richardson, K. K., Richardson, F. C., Crosby, R. M., Svenberg, J. A. & Skopek, T. R. *Proc. natn. Acad. Sci. U.S.A.* **84**, 344-348 (1987).
16. Goodlad, R. A. & Wright, N. A. in *Techniques in Life Sciences* Vol. 2 *Techniques in digestive physiology* (ed. Titchen, D. A.) 211-223 (Elsevier, Ireland, 1982).
17. Sater, W., Brennard, J., McMillan, S. & Fox, M. *Mutation Res.* **73**, 171-181 (1980).
18. Napalkov, N. P. & Alexandrov, V. A. *Z. Krebsforsch.* **71**, 32-50 (1968).
19. IARC. *IARC Monographs on the Evaluation of Carcinogenic Risk of Chemicals to Man* Vol. 1 (International Agency for Research on Cancer, 1972).
20. Russell, L. B. *Genetics* **92**, s153-s163 (1979).
21. Fahrig, R. & Neuhauser-Klaus, A. *J. Hered.* **76**, 421-426 (1985).

The Duchenne muscular dystrophy gene product is localized in sarcolemma of human skeletal muscle

Elizabeth E. Zubrzycka-Gaarn*, Dennis E. Bulman*, George Karpati†, Arthur H. M. Burghes*, Bonnie Belfall*, Henry J. Klamut*, Jim Talbot‡, Robert S. Hodges‡, Peter N. Ray* & Ronald G. Worton*

* Genetics Department and Research Institute, The Hospital for Sick Children and The Departments of Medical Genetics and Medical Biophysics, University of Toronto, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

† Neuromuscular Research Group, Montreal Neurological Institute, McGill University, Montreal, Quebec H3A 2BA, Canada

‡ Department of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada

Duchenne muscular dystrophy (DMD) and its milder form, Becker muscular dystrophy (BMD), are allelic X-linked muscle disorders in man¹. The gene responsible for the disease has been cloned from knowledge of its map location at band Xp21 on the short arm of the X chromosome²⁻⁵. The product of the DMD gene, a protein of relative molecular mass 400,000 (M_r 400K) recently named dystrophin, has been reported to co-purify with triads of mouse and rabbit skeletal muscle when assayed using polyclonal antibodies raised against fusion proteins encoded by regions of mouse DMD complementary DNA^{6,7}. Here we show that antibodies directed against synthetic peptides and fusion proteins derived from the N-terminal region of human DMD cDNA strongly react with an antigen present in skeletal muscle sarcolemma on cryostat sections of normal human muscle biopsies. This immunoreactivity is reduced or absent in muscle fibres from DMD patients but appears normal in muscle fibres from patients with other myopathic diseases. The same antibodies specifically react with a 400K protein in sodium dodecyl sulphate (SDS) extracts of normal human muscle subjected to Western blot analysis. We conclude that the product of the DMD gene is associated with the sarcolemma rather than with the triads and speculate that it strengthens the sarcolemma by anchoring elements of the internal cytoskeleton to the surface membrane.

We have recently reported the cloning of human adult muscle cDNA from the DXS206 locus and have confirmed that it is derived from the DMD/BMD gene⁵. To raise antibodies to the protein product of this gene and to identify the product in tissues, the sequence of a 2.0 kilobase (kb) cDNA was used to predict the amino-acid sequence of a portion of the protein product. Three synthetic peptides containing 11-13 amino acids derived from the N-terminal region of the protein were synthesized and conjugated to bovine serum albumin (BSA) or to keyhole limpet haemocyanin (KLH)⁸. The peptides do not carry any amino acid sequence homology with chicken α -actinin⁹ or spectrin¹⁰. Three fusion proteins were derived from a 1.8 kb cDNA fragment that was subcloned into pRIT and pATH fusion vectors^{11,12} from the internal *Bam*HI site to the end of the original 2.0 kb cDNA⁵.

As a first step in the identification of the protein product of the DMD gene, polyclonal antibodies were raised in rabbits to both the BSA-conjugated peptides and the fusion proteins. The titre of the antisera was determined using the solid phase enzyme-linked immunosorbent assay (ELISA) (ref. 13). Antibodies were purified by affinity chromatography on CNBr Sepharose 4B columns¹⁴. All the antibodies prepared against the synthetic peptides and fusion proteins generated similar results when used in experiments of the type described below. The data selected for presentation in this paper were obtained with the antibody P924 directed against the peptide PRFKSYAYTQA, and the antibody F927 raised against a 100K protein A-DMD fusion protein produced in the pRIT vector¹¹. This fusion protein contains ~60K of the DMD gene product (Fig. 1). Both antibodies were affinity purified and their specificity was analysed on Western blots of *Escherichia coli* extracts containing the Trp E-DMD fusion protein produced in a pATH vector¹². Antibodies made against both the synthetic peptide and the fusion protein reacted with the 100K fusion protein carrying the appropriate amino-acid sequence (Fig. 2a). This reactivity was not observed with the preimmune sera. The antibodies also precipitated the protein obtained from an *in vitro* translation of capped RNA prepared from the 2.0 kb cDNA (Fig. 2b). These results indicate that the antibodies recognize the protein product of the DMD/BMD gene.

Antibodies P924 and F927 were therefore used to detect the DMD protein and to estimate its size in normal human muscle. Protein from normal human skeletal muscle solubilized in SDS was fractionated according to molecular weight using a Sepha-

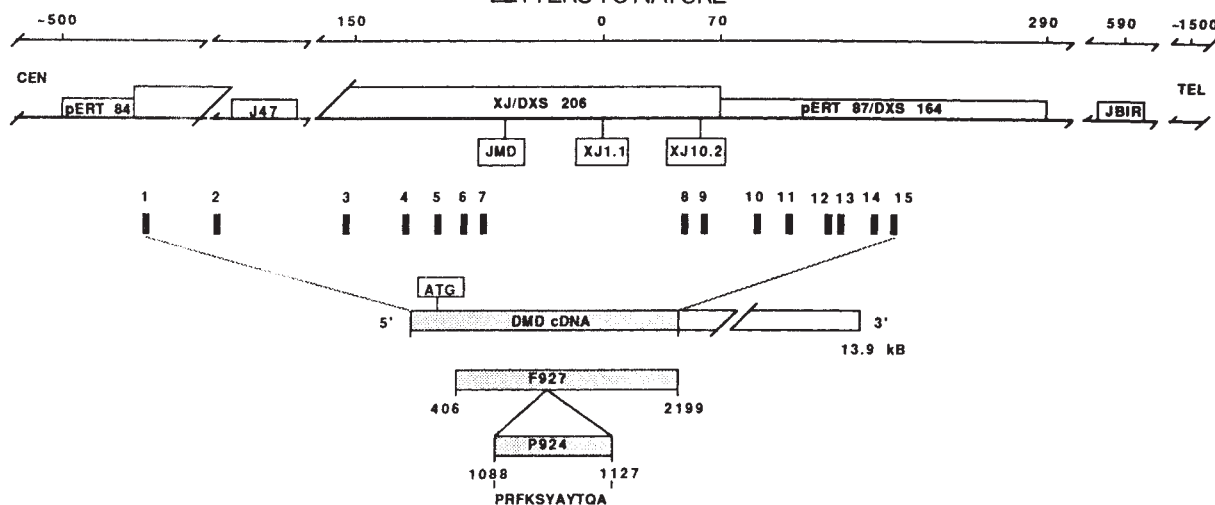


Fig. 1 Localization of cDNA sequences utilized to produce peptide and fusion protein antibodies relative to the DMD gene locus. A 2.0 kb cDNA clone⁵ containing the first 15 exons of the DMD gene was used to generate peptides and fusion proteins corresponding to the DMD gene product. The approximate locations of exons relative to cloned regions of the DMD locus are shown. The DMD locus spans approximately 2000 kb of band Xp21 on the short arm of the X-chromosome. Distances are indicated relative to the X:21 translocation break point (XJ1.1) used to clone the XJ region of the DMD gene locus²⁹. The position of XJ10.2, used to isolate the 2.0 kb cDNA, is shown. pERT and deletion jump clones described before³⁰ are also indicated. The nucleotide sequence corresponding to the 5' end of the DMD gene has been previously reported^{3,4}. The cDNA sequence used to generate antibody F927 spans the region from nucleotide 406–2199. The synthetic peptide (peptide 924) sequence corresponds to nucleotides 1088–1127. The ATG translational start site at position 208 is indicated.

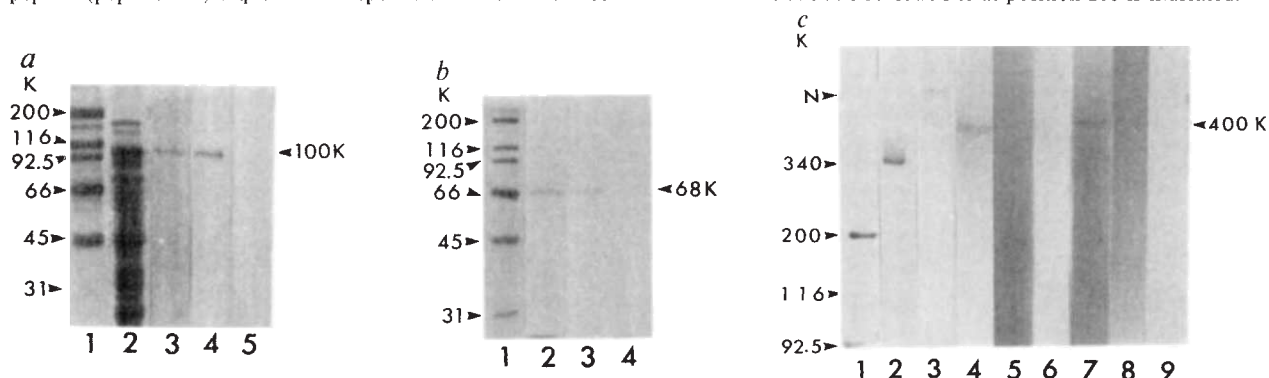


Fig. 2 Detection of the DMD gene protein product *in vitro* and in human muscle. *a*, Western blots of Trp E fusion protein. Lanes 1 and 2, molecular weight standards and *E. coli* extract, respectively, stained with Coomassie Brilliant Blue; Lanes 3–5, *E. coli* extracts immunostained with antibodies P924, F927 and preimmune serum, respectively. *b*, Immunoprecipitation of the *in vitro* translation product of the DMD gene. Lane 1, molecular weight standards; Lanes 2–4, autoradiogram of the translation product precipitated with P924, F927 and preimmune serum, respectively. *c*, Western blots of human muscle. Lanes 1 and 2, electroblotted α_2 macroglobulin (340K) and molecular weight standards, respectively, stained with Ponceau-S (Sigma); Lanes 3, 4 and 7, fractions eluted from a Sephacryl-S 500 column, containing muscle proteins of 200K and above, immunostained with polyclonal antibody to nebulin diluted 1:2000 (Lane 3), antibody P924 (Lane 4) or antibody F927 (Lane 7) both diluted 1:100; Lanes 5 and 8, fractions eluted from a Sephacryl-S500 column containing lower M_r muscle proteins immunostained with antibody P924 or F927, respectively; Lanes 6 and 9, total muscle extract immunostained with antibody P924 or F927, respectively.

Methods. *a*, A synthetic peptide (peptide 924) corresponding to 11 amino acids from the N-terminal region of the DMD gene (Fig. 1) was prepared by the Alberta Peptide Institute, Dept. of Biochemistry, University of Alberta, Edmonton, Alberta T6G 2H7, Canada. Trp E and protein A fusion proteins were prepared by cloning the 1.8 kb fragment of the original 2.0 kb cDNA from its internal Bam HI site to the end of the clone into pATH 11 and pRIT2T fusion vectors^{11,12}, respectively. The fusion proteins were induced in *E. coli* N4830-1 (Pharmacia) for the pRIT fusion and DH5 α (Bethesda Research Laboratories) for the pATH fusion. The protein A fusion protein (fusion protein 927) was purified from cell homogenates using an IgG Sepharose 6FF column (Pharmacia) followed by a formic acid cleavage of an aspartyl-proline peptide bond between protein A and the desired fusion protein³¹. Male New Zealand rabbits were immunized by subcutaneous injection with 400 μ g BSA-conjugated 924 peptide or with 300 μ g 927 fusion protein after purification on the IgG column but before cleavage with formic acid. In both cases the antigens were emulsified with Freund's complete adjuvant. Booster injections with the same amount of the antigens in incomplete Freund's adjuvant were given 2, 4 and 6 weeks after the first injection. Antisera were obtained 2 weeks after the first booster injection and several more times at 2 week intervals. Antibodies P924 and F927 were affinity purified¹⁴ using CNBr Sepharose 4B columns (Pharmacia) containing, respectively, KLH-conjugated peptide 924 and the 60K formic acid cleavage product of fusion protein 927 after removal of its protein A portion³¹. Total *E. coli* protein (10 μ g) containing Trp E fusion protein was separated in a 10% Laemmli gel³² and electroblotted onto nitrocellulose³³. The blot was blocked with 5% non-fat milk, incubated with primary antibody or preimmune serum (diluted 1:100) and then incubated with the peroxidase-conjugated goat antibody directed against rabbit IgG (Bio-Rad). The 2.0 kb cDNA was transcribed using T7 polymerase. The transcription reaction was carried out in the presence of the RNA cap structure m⁷G[5'ppp(5') Gm] (ref. 34). The capped RNA was then translated in a reticulocyte lysate system (Promega). The translation product (~68K) was precipitated with the antibody P924 or F927 using protein A-Sepharose 4B beads (Pharmacia) as described³⁵. The immunoprecipitates were separated in a 10% Laemmli gel³². The gel was stained with Coomassie Brilliant Blue, destained and exposed to Kodak X-Omat R film. *c*, Normal, adult muscle was minced and boiled in SDS-buffer¹⁶ and insoluble proteins were pelleted and discarded. Proteins present in the total muscle extract were fractionated according to their M_r using a superfine Sephacryl S-500 column (Pharmacia) as described¹⁵. The protein content of the samples was determined³⁶ and 50 μ g of the protein was separated in a 6% Laemmli gel³² with 1% crosslinker. The proteins were electroblotted according to Towbin³³ except that methanol was omitted from the transfer buffer. Immunostaining with the antibodies followed by incubation with the alkaline phosphatase-conjugated goat anti-rabbit second antibody (Promega Biotec) was performed using the Promega protocol.

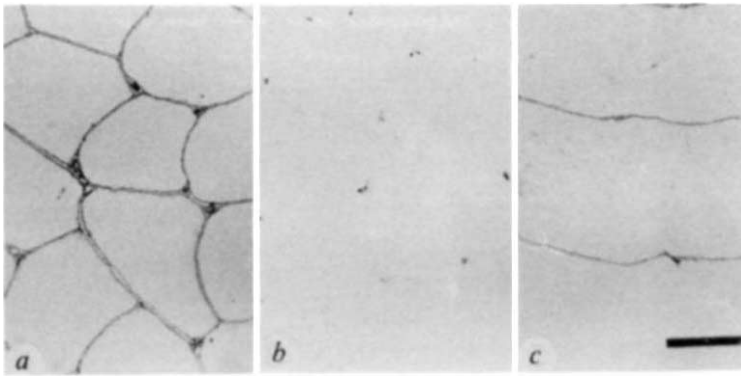
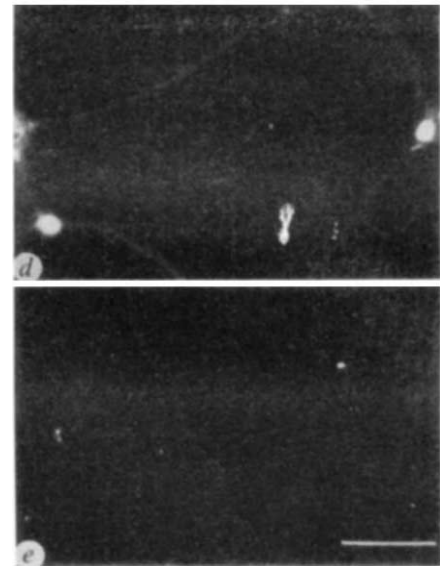


Fig. 3 Sections of normal human muscle fibres immunostained with F927 antibody. *a-c*, Peroxidase conjugated second antibody (scale bar: 50 μ m). *d* and *e*, Fluorescein-conjugated second antibody (scale bar, 25 μ m). *a, b, d, e*, transverse sections; *c*, longitudinal section; *a, c, d*, immune serum; *b, e*, preimmune serum.

Methods. Four μ m thick transverse or longitudinal cryostat sections were prepared from frozen blocks of muscle biopsies. The blocks were stored at -70°C . Freshly-cut sections were fixed for 60 s with acetone and washed with phosphate buffered saline (PBS). The primary antibody or preimmune serum diluted 1:20 was applied for 20 min at room temperature. Sections were then washed with PBS and incubated for 2 h at room temperature with the second antibody which alternatively was biotinylated anti-rabbit IgG followed by avidin-DH biotinylated horseradish peroxidase or anti-rabbit IgG conjugated to fluorescein. All panels in this figure were photographed and printed under constant conditions so the results with antibody and preimmune serum are directly comparable.



cryl S-500 column¹⁵. Fractions containing proteins of different M_r range were analysed by immunostaining on Western blots. The results indicate that both antibodies P924 and F927, recognize a protein band of $\sim 400\text{K}$ that migrates between nebulin and myosin (Fig. 2c, Lanes 4 and 7). Polyclonal antibodies prepared against nebulin (gift from Dr K. Maruyama, Chiba University) do not cross-react with this protein (Fig. 2c, Lane 3), showing that the DMD gene product is clearly distinct from nebulin as previously reported¹⁶. No bands were visible with preimmune sera (not shown). The affinity-purified antibodies also recognized a 400K protein in unfractionated extracts of normal human muscle, however the band was very faint (perhaps attributable to the fact that the antibodies were prepared against native protein) and not visible in photographic reproductions of the blot (Fig. 2c, Lanes 6 and 9). The whole muscle extract in Lanes 6 and 9 is shown to document that the antibodies do not react appreciably with other proteins present in the detergent extracts of normal human muscle. Staining of lower M_r proteins was also not seen in other fractions from the Sephacryl S-500 column (Fig. 2c, Lanes 5 and 8).

Microscopic localization of the DMD protein in human tissues was carried out using an indirect avidin-biotin based immunoperoxidase or immunofluorescence technique on cryostat sections of human muscle biopsies. The antibody directed against the fusion protein (F927) gave a consistent and strong staining or fluorescence of the sarcolemma of muscle fibres in sections of nine normal muscles (Fig. 3a,c,d). No staining or fluorescence was observed within muscle fibres. In particular, no intracellular staining of the triads was visualized, in contrast to positive staining of this structure obtained in our laboratory with several monoclonal and a polyclonal antibodies to the Ca^{2+} ATPase of the sarcotubular system¹⁷. Some weak staining of the endomysial space and of endomysial capillaries was seen. This staining was also present in specimens treated with the preimmune sera. The preimmune sera and all other technical controls, including a control with antibodies preabsorbed with purified fusion protein, showed no reactivity against sarcolemma (Fig. 3b and e). In six Duchenne specimens examined, sarcolemmal staining was absent or reduced in relation to normal controls (Fig. 5a). In sixteen disease controls other than DMD, the sarcolemmal staining appeared equal to that in normal muscle (Fig. 5b,c). Several other tissues were examined for

antibody staining. The surface membrane of myocardiocytes also exhibited faint staining but hepatocytes and smooth muscle cells from skin were negative.

The antibody directed against the peptide (P924) produced a similar but less intense sarcolemmal and endomysial staining in normal (Fig. 4) and non-DMD disease controls (not shown). No staining could be discerned in Duchenne muscle fibres. The stronger staining obtained with antibody F927 as compared to antibody P924 is probably due to the fact that antibody F927 recognizes many more epitopes. Another peptide antibody prepared to the synthetic peptide with amino-acid sequence AIEREKAERKFRKL derived from the original cDNA clone of Monaco *et al.*² also gave results similar to those obtained with P924 and F927. Antibodies P924 and F927 also gave a strong immunoperoxidase or immunofluorescence staining on the surface of normal human clonal myotubes in culture (manuscript in preparation), consistent with the sarcolemmal localization of the DMD gene product.

In conclusion, we have raised and characterized antibodies directed against synthetic peptides and fusion proteins whose amino-acid sequences were derived from the sequence of a

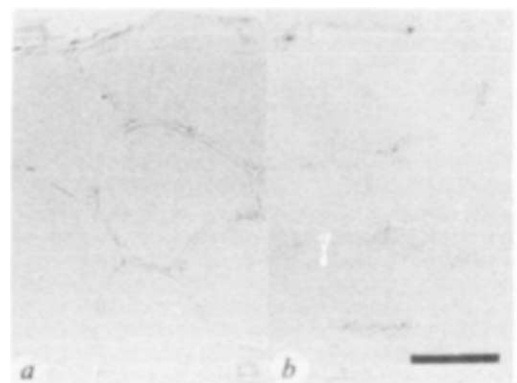
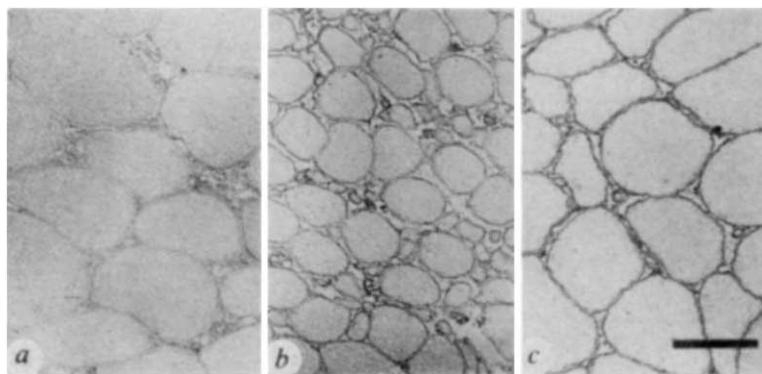


Fig. 4 Transverse sections of normal human muscle fibres immunostained with antibody P924. *a*, Normal muscle immunostained with antibody P924 diluted 1:20 as described in the legend to Fig. 3; *b*, the same muscle stained with the preimmune serum.

Fig. 5 Transverse sections of pathological muscle fibres immunostained with F927 antibody. *a*, 5-year-old DMD biopsy; *b*, 2-year-old nemaline myopathy biopsy; *c*, dermatomyositis biopsy. Sections were prepared and immunostained using peroxidase-conjugated second anti-rabbit as described in Fig. 3.



human adult muscle DMD cDNA recently cloned in our laboratory⁵. These antibodies react specifically with a 400K protein in normal human muscle. The size of the protein is in agreement with the size of the messenger RNA^{2,5,18} and in all likelihood it is the same protein (dystrophin) detected by Hoffman *et al.*^{6,7} using an antiserum against a fusion protein derived from a different DMD cDNA clone. The specific immunocytochemical reactivity observed in the sarcolemma of normal human muscle fibres with the antibodies F927 and P924 suggests that the DMD gene product is localized to the surface membrane of differentiated skeletal muscle fibres. The reduced or absent sarcolemmal reactivity in muscle of DMD patients is consistent with a defect in the DMD gene and with the observed genetic heterogeneity of Duchenne cases¹⁹. The heterogeneity may arise as a result of in-frame deletions which remove sequences having different degrees of functional significance.

At its N-terminus, the DMD gene product shows some homology to α -actinin and spectrin^{9,20}. However, the results presented here, are unlikely, for a number of reasons, to be due to any crossreactivity of the antibodies with these proteins. First, the antibodies used in this study do not recognize either of these proteins on Western blots. Furthermore, skeletal muscle α -actinin, a protein of ~100K, is localized to myofibrils²¹, and the sarcolemmal localization of the DMD gene product shown in this paper is therefore inconsistent with antibody cross-reactivity with α -actinin. Spectrin, a cytoskeletal protein of ~200K, is widely expressed in different tissues including skeletal muscle²². It has, however, been reported²³ that muscle fibres of patients with DMD show similar patterns of anti-spectrin antibody binding to that of normal individuals. The sarcolemmal localization of the DMD gene product described here is muscle specific and is not found in any of the non-muscle tissues studied. Moreover, the three synthetic peptides used for the antibody production do not show any homology to α -actinin and spectrin^{9,10}.

In contrast to our results, and those recently reported by Sugita *et al.*²⁴, subcellular muscle fractionation has suggested an intracellular localization of dystrophin to triads in mouse and rabbit skeletal muscle⁷. The results presented in this paper and those described by Hoffman *et al.*⁷ have been obtained by methods which are not easily comparable. The cofractionation of sarcolemmal vesicles with triadic structures should be considered, however, especially in view of the fact that the transverse tubular component of triads represents intracellular invaginations of the sarcolemma.

The sarcolemmal localization of dystrophin is consistent with the 'membrane theory' of the disease^{25,26}. Electron microscopic studies of Duchenne muscle have indicated that in non-necrotic muscle fibres of DMD patients, the plasma membrane is often found locally separated from the basal lamina²⁶. If such localized separations are responsible for the focal lysis of plasma membrane precipitating segmental necrosis, one might suggest that dystrophin is a binding molecule that anchors elements of the internal cytoskeleton to the plasma membrane thereby providing mechanical strength for the sarcolemma. According to this scheme, the function of dystrophin may be analogous to integrins²⁷. Without normal dystrophin, large diameter muscle fibres would be particularly susceptible to contraction-induced plasma membrane tearing and muscle necrosis²⁸. This model would be consistent with many known facts about the disease pathology^{25,26}.

We thank Dr K. Maruyama, Chiba University, Chiba, Japan for his kind gift of polyclonal antibodies against nebulin and N. Kasmierski for her help in the preparation of the manuscript. This work was supported by grants from the Medical Research Council of Canada, the Muscular Dystrophy Association of Canada, the Muscular Dystrophy Association, U.S.A., and the Killam Professional Chair Endowment (G.K.) of the Montreal Neurological Institute. H.J.K. is a recipient of an MDA, USA postdoctoral fellowship award.

Received 17 March; accepted 5 May 1988.

- Worton, R. G. & Burghes, A. H. M. *Int. Rev. Neurol.* **29**, 1-75 (1988).
- Monaco, A. P. *et al. Nature* **323**, 646-650 (1986).
- Koenig, M. *et al. Cell* **50**, 503-517 (1987).
- Hoffman, E. P., Monaco, A. P., Feener, C. C. & Kunkel, L. M. *Science* **238**, 347-350 (1987).
- Burghes, A. H. M. *et al. Nature* **328**, 434-437 (1987).
- Hoffman, E. P., Brown, R. H. & Kunkel, L. M. *Cell* **51**, 919-928 (1987).
- Hoffman, E. P., Knudson, C. M., Campbell, K. P. & Kunkel, L. M. *Nature* **330**, 754-758 (1987).
- Parker, J. M. R. & Hodges, R. S. J. *Protein Chem.* **3**, 465-478 (1985).
- Hammonds, R. G. *Cell* **51** (1987).
- Lipman, D. J. & Pearson, W. R. *Science* **227**, 1435-1441.
- Nilsson, B., Abrahamson, L. & Uhlén, M. *EMBO J.* **4**, 1075-1080 (1985).
- Spindler, K. R., Rossen, D. S. E. & Berk, A. J. *J. Virol.* **49**, 132-141 (1984).
- Engvall, E. & Perlmann, P. *Immunochemistry* **8**, 871-879 (1971).
- Affinity Chromatography: Principles and Methods* 8-15 (Pharmacia Fine Chemicals, Sweden, 1977).
- Wang, K., McClure, J. & Tu, A. *Proc. natn. Acad. Sci. U.S.A.* **76**, 3698-3702 (1979).
- Sugita, H. *et al. Proc. Japan Acad.* **63**, 107-110 (1987).
- Karpati, G. *et al. Ann. Neurol.* **20**, 38-49 (1986).

- Lev, M. A., Feener, C. C., Kunkel, L. M. & Brown, R. H. *J. Biol. Chem.* **262**, 15817-15820 (1987).
- Hyser, C., Province, M. & Griggs, R. C. *Ann. Neurol.* **22**, 553-555 (1987).
- Davison, M. D. & Critchley, D. R. *Cell* **52**, 159-160 (1988).
- Ebashi, S., Ebashi, F. & Manuyama, K. *Nature* **203**, 645-646 (1964).
- Lazarides, E. & Nelson, W. J. *Cell* **31**, 505-508 (1982).
- Appleyard, S. T. *et al. Proc. natn. Acad. Sci. U.S.A.* **81**, 776-780 (1984).
- Sugita, H. *et al. Proc. Japan Acad.* **64**, 37-39 (1-988).
- Mokri, B. & Engel, A. G. *Neurology* **25**, 1111-1120 (1975).
- Carpenter, S. & Karpati, G. *Brain* **102**, 147-161 (1979).
- Ruoslatti, E. & Pierschbacher, M. D. *Science* **238**, 491-497 (1987).
- Karpati, G. & Carpenter, S. *Am. J. med. Genet.* **25**, 653-658 (1986).
- Ray, P. N. *et al. Nature* **318**, 671-675 (1985).
- Kunkel, L. M. *et al. Proc. natn. Acad. Sci. U.S.A.* **82**, 4778-4782 (1985).
- Piszkiewicz, D., Landon, M. & Smith, E. L. *Biochem. biophys. Res. Commun.* **40**, 1173-1178 (1970).
- Laemmli, U.K. *Nature* **227**, 680-685 (1970).
- Towbin, H., Stachelin, T. & Gordon, J. *Proc. natn. Acad. Sci. U.S.A.* **76**, 4350-4354 (1979).
- Nielsen, D. A. & Shapiro, D. J. *Nucleic Acids Res.* **14**, 5936 (1986).
- Zubrzycka-Gaarn, E. E. *et al. J. biol. Chem.* **258**, 4576-4581 (1983).
- Lowry, D. H. *et al. J. biol. Chem.* **193**, 265-275 (1951).

Molecular cloning and expression of the major protein kinase C substrate of platelets

Michael Tyers*, Richard A. Rachubinski*, Mary I. Stewart†, Angela M. Varrichio‡, Robert G. L. Shorr‡, Richard J. Haslam*† & Calvin B. Harley*

* Department of Biochemistry and † Department of Pathology, McMaster University, Hamilton, Ontario, Canada L8N 3Z5

‡ Smith Kline and French Laboratories, PO Box 1539, King of Prussia, Pennsylvania 19406, USA

In platelets, agonists that stimulate phosphoinositide turnover cause the rapid phosphorylation of a protein of apparent relative molecular mass (M_r) 40–47,000, called P47, by protein kinase C (PKC)^{1–4}. Diverse identities have been ascribed to P47 including lipocortin⁵, inositol 1,4,5-trisphosphate 5-phosphomonoesterase⁶, pyruvate dehydrogenase α subunit⁷ and an actin regulatory protein⁸. We have isolated human P47 clones by immunological screening of a λ gt11 complementary DNA library from HL-60 cells, a human promyelocytic leukaemia cell line⁹. P47 recombinants thus identified hybridized to a 3.0 kilobase (kb) messenger RNA in mature white blood cell lines; the same mRNA was induced in HL-60 cells during differentiation. A 1,050 base pair (bp) open reading frame that could encode a protein of M_r 40,087 was confirmed by comparison with peptide sequences from platelet P47, and by expression of the putative recombinant P47 in *E. coli* and *in vitro*. The P47 sequence appears to have been conserved throughout vertebrate evolution, and is not similar to any other known sequence including human lipocortin and the α subunit of pyruvate dehydrogenase. The P47 protein contains a potential Ca^{2+} -binding 'EF-hand' structure and a region that strongly resembles known PKC phosphorylation sites.

Platelets have been widely used to investigate mechanisms of signal transduction, particularly those involving hydrolysis of phosphatidylinositol 4,5-bisphosphate in which the second messengers diacylglycerol and inositol 1,4,5-trisphosphate respectively activate PKC and release Ca^{2+} from intracellular stores¹. Platelet agonists such as thrombin trigger a marked phosphorylation of P47 by PKC which usually precedes secretion of platelet granule contents^{1–4}. Several roles have been suggested for P47 in stimulus-response coupling, including that it is a member of the lipocortin family⁵; that phosphorylated P47 attenuates Ca^{2+} release from intracellular stores by degrading inositol 1,4,5-trisphosphate⁶; that it is in fact the α subunit of pyruvate dehydrogenase⁷; and, more recently, that non-phosphorylated P47 inhibits actin polymerization, implicating it in control of cytoskeletal reorganization during secretion⁸. To facilitate the study of P47 structure and function, we undertook to clone P47 from HL-60 cells, a human promyelocytic leukaemia cell line⁹ that becomes enriched for P47 when induced to differentiate with retinoic acid¹⁰.

A rabbit antiserum to human P47 (M.I.S., Lynham, J. A., Sartori, C. S., Davidson, M. M. L., Culty, M. & R.J.H., manuscript in preparation) was used to screen a λ gt11 cDNA library constructed from HL-60 cells that had been treated with retinoic acid for 5 days. Immunoreactive clones were obtained at the expected frequency based on P47 protein abundance¹⁰ and were shown to encode the P47 protein sequence (see below). All clones contained inserts that hybridized to a 3.0 kb mRNA regulated during retinoic acid-induced HL-60 differentiation (Fig. 1a) in a similar fashion to P47 protein¹⁰. In addition, the P47 mRNA levels were transiently elevated in response to 12-*O*-tetradecanoyl-phorbol-13-acetate (TPA, Fig. 1a) suggesting that the P47 gene may contain phorbol ester responsive elements¹¹

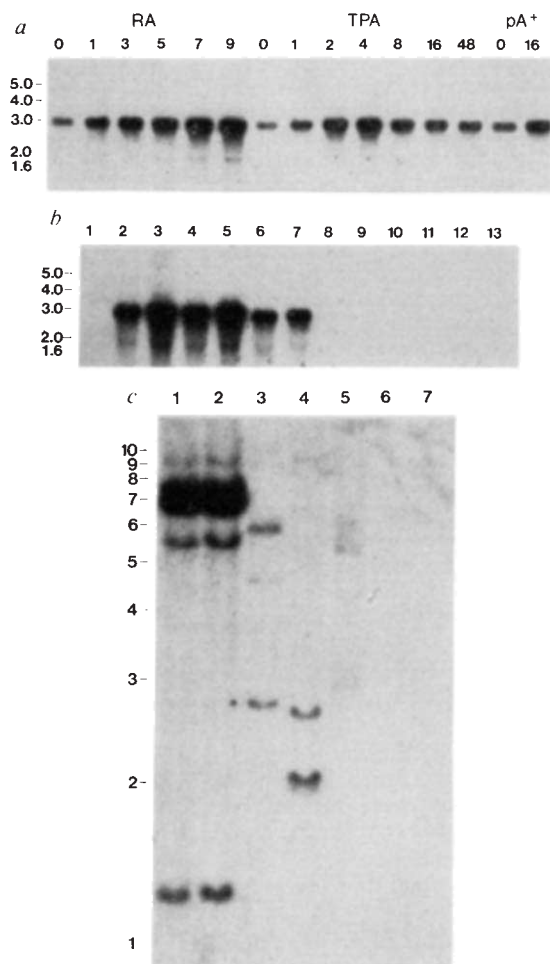
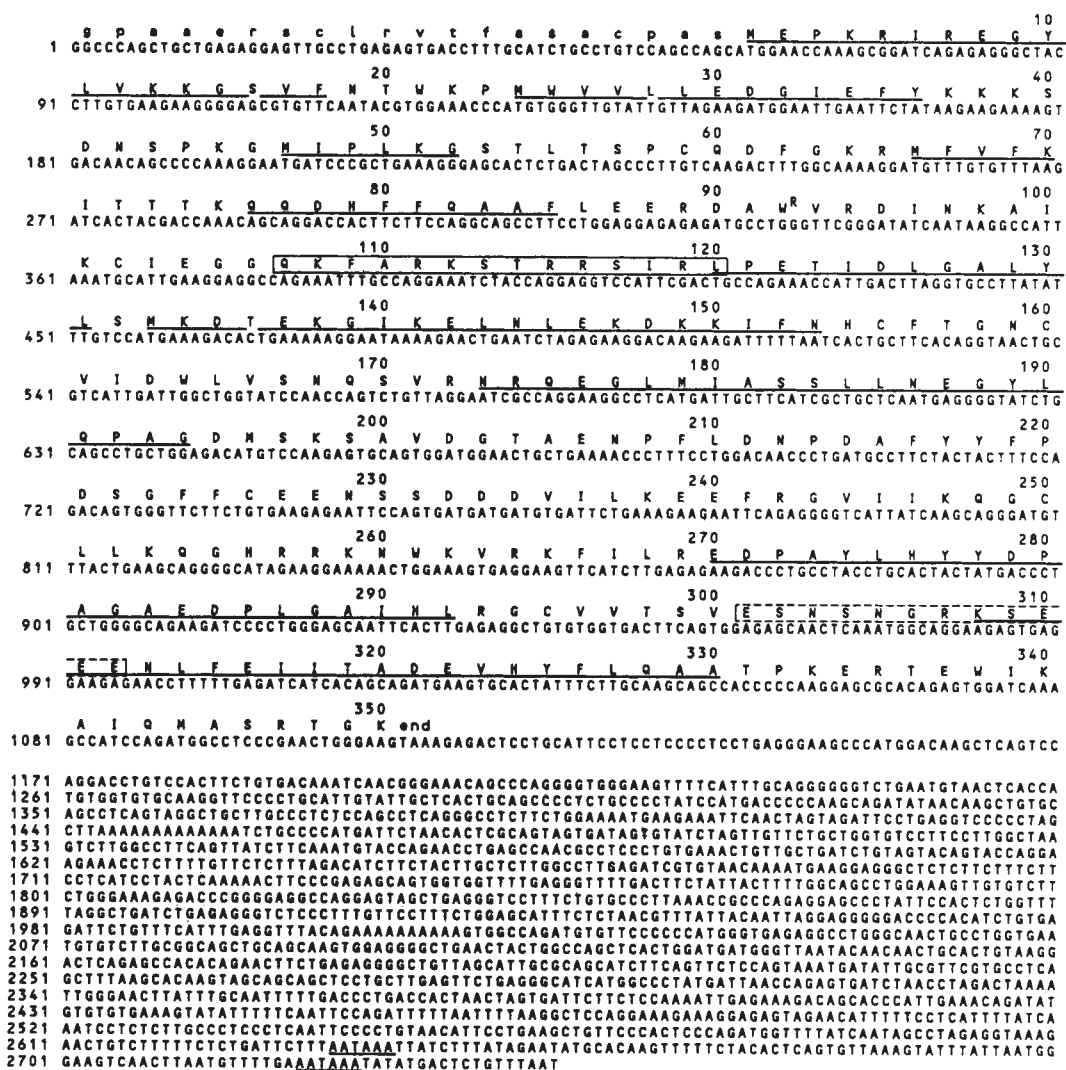


Fig. 1 *a*, Regulation of P47 mRNA during HL-60 cell differentiation. Northern analysis of total RNA from cells treated with 1 μ M retinoic acid (RA) for the indicated number of days or 20 nM TPA for the indicated number of hours, and of poly(A)⁺ RNA from cells treated with TPA for 0 and 16 hours are shown. *b*, Tissue distribution of P47 mRNA. Human cell lines were used as sources of tissue specific RNA: K562 undifferentiated myeloblastic leukaemia (lane 1), HL-60 promyelocytic leukaemia (lane 2), U937 histiocytic leukaemia (lane 3), THP-1 monocytic leukaemia (lane 4), Raji Burkitts lymphoma (lane 5), JO5 Epstein-Barr virus transformed B cell (lane 6), HUT 78 T cell (lane 7), HepG2 hepatoma (lane 8), MJ fibroblast (lane 9), COLO 16 keratinocyte (lane 10), CEN lens epithelial (lane 11), SH neuroblastoma (lane 12), SH neuroblastoma differentiated with 10 ng ml⁻¹ nerve growth factor for 48 h (lane 13). *c*, Southern analysis of the P47 gene in various species: human (HL-60 cells, lane 1; normal leukocytes, lane 2), mouse (lane 3), chicken (lane 4), rainbow trout (lane 5), *Drosophila melanogaster* (lane 6), *Saccharomyces cerevisiae* (lane 7). Marker sizes in kb or kbp are indicated to the left of each blot. **Methods.** *a*, *b*, RNA was isolated as described²⁷, electrophoresed (15 μ g total, 0.5 μ g poly(A)⁺) on a 1% agarose, 6% formaldehyde gel, transferred to nitrocellulose and probed with 10⁷ c.p.m. of hexamer-primed P47 cDNA *Pvu*II fragment (nt 8–2085). Hybridization was in 50% formamide, 5 $\times$ Denhardt's solution, 5 $\times$ SSC, 0.1% SDS, 50 mM NaH₂PO₄, 5 mM Na₂H₂P₂O₇ and 250 μ g ml⁻¹ salmon sperm DNA. Blots were washed with 2 $\times$ SSC, 0.1% SDS at room temperature then 0.1 $\times$ SSC, 0.1% SDS at 48 $^{\circ}$ C. *c*, 10 μ g of DNA from each species was digested with *Eco*RI then electrophoresed on 0.8% agarose, denatured, transferred to nitrocellulose and probed as above except that hybridization was in 30% formamide and the high temperature wash was in 0.5 $\times$ SSC.

Fig. 2 Nucleotide and deduced amino acid sequence of P47 cDNA. The 350 amino acids of P47 encoded by the 1,050 nt open reading frame are shown in upper case one-letter code; the 5' extension of this reading frame is shown in lower case. Peptide sequences from platelet P47 are indicated by underlines interrupted by gaps where residues could not be unequivocally identified. Possible polyadenylation signals are underlined. A strong candidate region for phosphorylation by PKC is boxed by solid lines and the Ca²⁺-binding loop of the putative EF-hand is boxed by broken lines.

Methods. Poly(A)⁺ RNA from HL-60 cells treated with retinoic acid for 5 d was used to construct cDNA as described²⁸ except that the double stranded cDNA hairpin was blunt ended with 40 U mung bean nuclease μg^{-1} cDNA. Phosphorylated *Eco*RI linkers (300 ng) were ligated to 1 μg cDNA and, after digestion with *Eco*RI, removed by passage over a Sepharose CL-4B column. Linkered cDNA (0.5 μg) was ligated to dephosphorylated λ gt11 arms and packaged *in vitro* (Promega Biotech). One tenth of the unamplified library (3×10^5 plaques), was screened with a rabbit polyclonal P47 antiserum (M.I.S. *et al.*, manuscript in preparation) and goat-anti-rabbit alkaline phosphatase conjugate according to the supplier (Promega Biotech). Inserts from immunoreactive clones were subcloned into pUC 118/119 and sequenced in both directions by dideoxy chain termination using specific primers and nested sets of deletions²⁹. The sequence of the longest clone (nt 1-2724) is shown with 21 nt of 3' sequence appended from other inserts. A single discrepant residue present in two of five different clones in the coding region is indicated in superscript at amino acid position 92. To confirm the P47 amino-acid sequence, human platelet P47, purified to near homogeneity by successive column chromatography on DEAE-cellulose, DEAE-sephadex, phenyl-sepharose and SE-sepharose (ref. 13 and M.I.S. *et al.*, manuscript in preparation) was denatured, alkylated, and further purified by reverse phase HPLC on a 330 A Vydac C₄ column using a linear gradient of 0-70% acetonitrile in aqueous 0.1% trifluoroacetate. Peptides were generated by cyanogen bromide or tryptic digestion. The former were resolved on the same column as for final purification of P47, and the latter on a Beckman C₁₈ column. A Beckman 890 M sequencer was used and resultant phenylthiohydantoin-amino acids were analysed by reverse phase HPLC on a C₁₈ column as described³⁰.



which could modulate its expression according to the level of stimulation of PKC. The tissue distribution of the P47 mRNA was investigated using human cell lines to avoid contamination from platelets and leukocytes in which P47 is expressed. P47 transcripts were present in haemopoietic cells of myeloid and lymphoid origin but not in K562 cells, a more primitive line¹², nor in lines derived from liver, fibroblasts, skin, lens or nerve tissue (Fig. 1b). This result corroborates immunological data (M.I.S. *et al.*, manuscript in preparation) and, together with the increased expression in differentiated HL-60 cells, indicates that P47 may have a common function in platelets and leukocytes. Southern blot analysis identified a single locus encompassing ~30 kb of the human genome that hybridized to the cDNA clone (Fig. 1c). The restriction map was the same in HL-60 cells and in normal human tissue, indicating that no major rearrangements occurred in the P47 gene during genesis of the HL-60 line. The gene appeared to be conserved in vertebrates at least as well as *c-myc* (data not shown) but could not be detected in insects or yeast (Fig. 1c). In conjunction with the

tissue distribution data, this suggests that P47 plays an important role in highly developed haemopoietic systems.

The longest cDNA inserts (2.75 kb) were sequenced and found to contain a 5' unbounded open reading frame of 1,110 nucleotides (nt) that could code for a protein of M_r 42,064 (Fig. 2), somewhat smaller than the apparent M_r of P47 (ref. 13). The N-terminus of purified human platelet P47 could not be determined by protein sequencing because the α -amino group was blocked. Instead, to confirm the identity of the cDNA clones, cyanogen bromide and tryptic peptides from P47 were sequenced. In all positions for which unequivocal peptide sequence was obtained, the amino-acid sequence deduced from the cDNA sequence was identical (Fig. 2). Although the longest clones were nearly full length as judged by the size of P47 mRNA detected in cell lines, it was possible that some 5' coding sequence was missing. However, comparison of nucleotides flanking the first ATG with the consensus for initiation of translation¹⁴, revealed a near perfect match (CCAGCATGG versus CCACCATGG). This putative initiation site was confirmed by

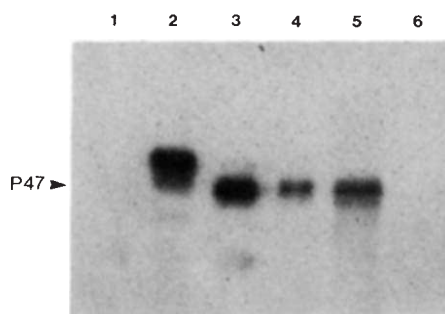


Fig. 3 Expression of the P47 coding region in *E. coli* and *in vitro*. Lanes 1-4, Immunoblots of P47 from expression of cDNA fragments in *E. coli*: lane 1, no insert; lane 2, nt 8-1152 corresponding to a protein of M_r 42,112; lane 3, nt 61-1152 corresponding to a protein of M_r 40,087; lane 4, 100 ng platelet P47. Lanes 5 and 6, *in vitro* transcription and translation of P47 cDNA: lane 5, nt 8-2085; lane 6, no insert control.

Methods. The *PvuII*-*NcoI* fragment of P47 cDNA (nt 8-1152) was inserted into the *NcoI* site of pKK-233 using *NcoI* linkers to create a vector expressing the complete open reading frame. Including two amino acids contributed by the linker sequence, this corresponded to a protein of M_r 42,112. In another construct, the G at -2 with respect to the putative initiator ATG (nt 59) was mutated³¹ using a mismatch oligonucleotide 5'-GTTCCATGGTGGCTGGA-3'. This generated an *NcoI* site at the first methionine which allowed precise cloning of the P47 coding region into the translation initiation site of pKK-233. *E. coli* JM109 containing each construct were grown to saturation overnight and lysed in sample buffer. Authentic size P47 (lane 3) accumulated to 25% of bacterial protein in the form of inclusion bodies. Samples were electrophoresed on an SDS-polyacrylamide gel (11%), transferred to nitrocellulose and reacted¹⁰ with P47 antiserum in the presence of control JM109 bacterial lysate. For *in vitro* transcription and translation, the *PvuII* fragment of P47 cDNA (nt 8-2085) was cloned into the *SmaI* site of pGEM3Z and reactions were carried out according to the supplier (Promega Biotech). Translations (1 μ l of each, corresponding to 3×10^5 c.p.m. from the P47 template) were electrophoresed on the same gel as immunoblot samples; as only 3 ng of *in vitro* synthesized P47 was loaded, the autoradiographic signal arises from ³⁵S-methionine, not ¹²⁵I-labelled protein A.

in vitro transcription and translation of P47 cDNA, the product of which co-migrated with purified platelet P47 (Fig. 3). To ensure that full size P47 was not generated by fortuitous initiation *in vitro*, the ATG flanking sequence was mutated to create an *NcoI* site (and also a consensus translation initiation signal) that allowed precise expression of the P47 coding region in *E. coli*. The recombinant protein co-migrated with P47 whereas a separate construct engineered to contain the entire open reading frame, including that coding for 20 amino acids 5' to the first ATG, gave a product with an apparent M_r 2,600 larger than platelet P47 (Fig. 3). As the most N-terminal CNBr peptide from platelet P47 began with the second residue of the cloned P47, it was the initiator methionine itself that was blocked, probably by acetylation¹⁵.

The complete P47 coding sequence thus specifies a protein of M_r 40,087. A T to C transition at nt 334 caused a tryptophan to arginine substitution in two of five independent clones sequenced (Fig. 2a). This sequence heterogeneity may in part account for the P47 isoforms seen on two-dimensional gels¹³. P47 mRNA has a long 3' untranslated region containing two poly(A) tracts, one of which (nt 1445-1457) was surrounded by several single base changes amongst different clones. The extreme 3' end of P47 mRNA has two possible polyadenylation signals¹⁶ but none of the clones contained a polyadenylated tail. Analysis of the deduced amino acid sequence of P47 shows basic peaks at both N and C termini, which, if protease-sensitive, could account for P42, a related protein that copurifies with P47 (M.I.S. *et al.*, manuscript in preparation). P47 has a high predicted α -helical

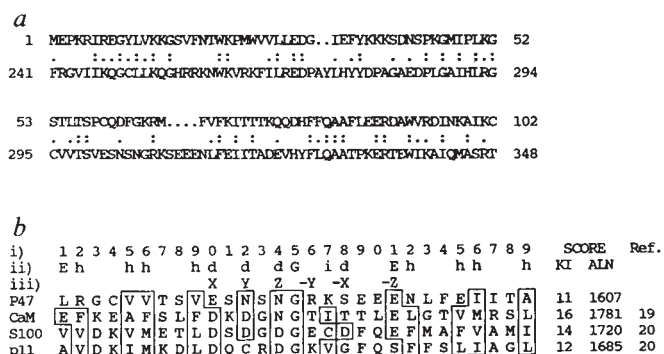


Fig. 4 a, Duplicated region in P47 found by alignment of P47 against itself. The alignment score of these regions was eight standard deviations greater than the mean alignment score of corresponding random P47 subsequences. b, Comparison of predicted P47 EF-hand to similar structures in known Ca^{2+} -binding proteins: i) numbering system proposed by Kretsinger¹⁷; ii) consensus sequence¹⁷ ($h = A, F, I, L, M, V$ or Y ; $d = D, E, N, Q, S$, or T ; $i = I$ or V); iii) octahedral vertices formed by liganding oxygens; P47 (292-320); CaM, calmodulin (19-37); S100 (β 52-80); P11 (50-78). The degree of similarity, based on the criteria of Kretsinger¹⁷ (KI) and a modified multiple sequence alignment program¹⁸ (ALN) in which only the 16 conserved residues are aligned against a core set of functional EF-hands, is shown by scores on the right.

content (37%) and is highly polar, consistent with its cytosolic location in platelets and HL-60 cells (data not shown). In addition there is significant internal homology between residues 1-102 and 241-348 (Fig. 4a), suggesting evolution by gene duplication. P47 contains a potential Ca^{2+} binding EF-hand motif¹⁷ at the C-terminus (Fig. 2). Comparison¹⁸ to functional EF-hands such as those in calmodulin¹⁹ and S100 (ref. 20) revealed that the Ca^{2+} -liganding residues are conserved, but that an anomalous lysine in the middle of the Ca^{2+} binding loop and non-matching residues in the putative E and F helices make it uncertain if P47 can bind Ca^{2+} (Fig. 4b). In contrast, p11, which scores higher than P47, is missing liganding residues and does not bind Ca^{2+} (ref. 20). P47 contains potential sites for phosphorylation by protein kinase C, characterized by serine and threonine residues adjacent to basic residues²¹⁻²⁴. Comparison¹⁸ with known substrates of protein kinase C identified amino-acid residues 107-120 of P47 as the most probable region of phosphorylations (Fig. 2). This sequence (QKFKARSTRSIRL) closely resembles the pseudosubstrate site of PKC (RFARKGSLRQ) (ref. 22). It has three probable phosphorylation sites, perhaps analogous to the phosphorylated regions of myelin basic protein²³, and myosin light chain²⁴. There are also other potential phosphorylation sites in P47 at serine 40, 43 and threonine 73.

Searches of current databases did not reveal any striking similarities of P47 to known sequences. In particular, P47 was not similar to the human lipocortin sequence²⁵ or the human pyruvate dehydrogenase α subunit sequence²⁶, proving that it is neither of these proteins. The P47 sequence yields no further direct clues as to other possible functions. Other work has shown that P47 and platelet inositol 1,4,5-trisphosphate 5-phosphomonoesterase can be partially separated (M.I.S. *et al.*, manuscript in preparation). We are now testing both human platelet P47 and recombinant P47 for Ca^{2+} -binding and actin-regulating activities. As P47 is unrelated to known proteins and its true M_r , structure, and cellular distribution have now been determined, we propose that it should be renamed pleckstrin (platelet and leukocyte C kinase substrate; also KFARKSTRSIR, the most probable phosphorylation site).

We thank Dr A. B. Futcher for discussion, Vivian Li and Eric Harley for help with the ALIGN program, and Gillian Regoecci,

Claudio Sartori and Mark Strohsacker for technical assistance. Database searches and sequence analyses used the NBRF Protein Identification Resource (Georgetown, MD) and the University of Wisconsin GCG software (processed on the National Research Council CND/SND system, Ottawa, Ontario).

Received 7 March; accepted 14 April 1988.

- Nishizuka, Y. *Nature* **308**, 693-698 (1984).
- Lyons, R. M., Stanford, N. & Majerus, P. W. *J. clin. Invest.* **56**, 924-936 (1975).
- Haslam, R. J. & Lynham, J. A. *Biochem. biophys. Res. Commun.* **77**, 714-722 (1977).
- Haslam, R. J., Lynham, J. A. & Fox, J. E. B. *Biochem. J.* **178**, 397-406 (1979).
- Touqui, L. *et al. Nature* **321**, 177-180 (1986).
- Connolly, T. M., Lawing, W. J. Jr. & Majerus, P. W. *Cell* **46**, 951-958 (1986).
- Chiang, T. M., Kang, E. S. & Kang, A. H. *Archs Biochem. Biophys.* **252**, 15-23 (1987).
- Hashimoto, K. *et al. Biochem. International* **14**, 759-767 (1987).
- Collins, S. J. *Blood* **70**, 1233-1244 (1987).
- Tyers, M., Rachubinski, R. A., Sartori, C. S., Harley, C. B. & Haslam, R. J. *Biochem. J.* **243**, 249-253 (1987).
- Lee, W., Mitchell, P. & Tjian, R. *Cell* **47**, 741-752 (1987).
- Clarke, B. J., Liao, S., Leeds, C., Soamboonsrup, P. & Neame, P. B. *Cancer Res.* **47**, 4254-4259 (1987).
- Imaoka, T., Lynham, J. A. & Haslam, R. J. *J. biol. Chem.* **258**, 11404-11414 (1983).
- Kozak, M. *Nucleic Acids Res.* **15**, 8125-8149 (1987).
- Augen, J. & Wold, F. *Trends biochem. Sci.* **11**, 494-497 (1986).
- Proudfoot, N. J. & Brownless, G. G. *Nature* **263**, 211-214 (1976).
- Kretsinger, R. H. *Ann. N.Y. Acad. Sci.* **356**, 14-19 (1980).
- Bacon, D. & Anderson, W. F. *J. molec. Biol.* **191**, 153-161 (1986).
- Waterson, D. M., Sharief, F. & Vanaman, T. C. *J. biol. Chem.* **255**, 962-975 (1980).
- Gerke, V. & Weber, K. *EMBO J.* **4**, 2917-2920 (1985).
- Woodgett, J. R., Gould, K. L. & Hunter, T. *Eur. J. Biochem.* **161**, 177-184 (1986).
- House, C. & Kemp, B. E. *Science* **238**, 1726-1728 (1987).
- Turner, R. S., Kemp, B. E., Su, H. & Kuo, J. F. *J. biol. Chem.* **260**, 11503-11507 (1985).
- Ikebe, M., Hartshorne, D. J. & Elzinga, M. *J. biol. Chem.* **262**, 9569-9573 (1987).
- Wallner, B. P. *et al. Nature* **320**, 77-81 (1986).
- Koike, K., Ohta, S., Urata, Y., Kagawa, Y. & Koike, M. *Proc. natn. Acad. Sci. U.S.A.* **85**, 41-45 (1988).
- Harley, C. B. *Gene anal. Tech.* **4**, 17-22 (1987).
- Wickens, M. P., Buell, G. N. & Schimke, R. T. *J. biol. Chem.* **253**, 2483-2495 (1978).
- Henikoff, S. *Gene* **28**, 351-359 (1984).
- Hawke, D., Yuan, P. & Shively, J. E. *Analyt. Biochem.* **120**, 302-311 (1982).
- Zoller, M. J. & Smith, M. *DNA* **3**, 479-488 (1984).

An *in vivo* recombinant RNA capable of autocatalytic synthesis by Q β replicase

Alexander V. Munishkin, Leonid A. Voronin
& Alexander B. Chetverin*

Institute of Protein Research, Academy of Sciences of the USSR,
142292 Pushchino, Moscow Region, USSR

A variety of small RNAs ranging from tens to hundreds of nucleotides in length grow autocatalytically in a Q β replicase (Q β phage RNA-dependent RNA polymerase) reaction in the absence of added template, and similar RNAs are found in Q β phage-infected *Escherichia coli* cells¹⁻⁴. Three such RNAs have been sequenced⁵⁻⁷. One of them that is 221 nucleotides (nt) long ('MDV-1' RNA) has been found to be partially homologous to Q β phage RNA⁸, which might be considered as an indication of its origination from by-products of the Q β RNA replication. To gain further insight into the origin and function of these RNAs, we have sequenced a new RNA, 120 nt long, isolated from the products of spontaneous synthesis by the nominally RNA-free Q β replicase preparation. The minus strand of this RNA appeared to be a recombinant RNA, composed of the internal fragment of Q β RNA (~80 nt long) and the 33-nt-long 3'-terminal fragment of *E. coli* tRNA^{Asp}. This seems to be the first strong indication of RNA recombination in bacterial cells. The various implications of this finding are discussed.

Figure 1a shows an electrophoretic pattern of the RNAs grown spontaneously in a purified Q β replicase preparation. The RNA comprising about half of the material synthesized was separated into individual complementary strands (Fig. 1b), sequenced and denoted as RQ120 RNA (that is, replicated by Q β replicase, 120 nt long).

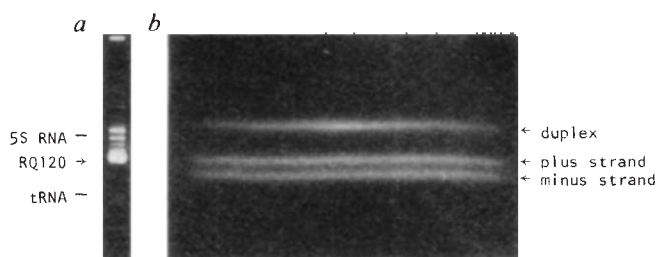


Fig. 1 Electrophoretic separation of the products of spontaneous RNA synthesis in cell-free Q β replicase system. Purified Q β replicase preparation lacking nucleotide absorption in the ultraviolet was isolated by a modified procedure of Blumenthal⁹ from the Q β amb86 phage-infected *E. coli* cells. *a*, RNAs synthesized by the pure enzyme in the presence of all four nucleoside triphosphates (0.4 mM each) under conditions of Mills *et al.*⁶ were separated in 10% polyacrylamide gel under denaturing conditions (7 M urea, 2 M EDTA, 40 °C) and stained with ethidium bromide. The most prominent RNA (RQ120) was isolated, amplified, 3'-labelled with [³²P]pCp by T4 RNA ligase and separated into individual complementary strands by electrophoresis (*b*) in a 10% gel in the presence of 1 mM MgCl₂ (a negative from the autoradiogram is shown). The slower migrating strand was designated as plus one as its content was usually 5-10% higher¹.

The secondary structure model of the RQ120 plus strand compatible with the ribonuclease digestion data consists of classical hairpins, the terminal (5')GGGA... and ...UCCCA(3') stretches being hydrogen-bonded to form a 'stalk' of the RNA molecule (Fig. 2a). From its resistance to a single strand-specific ribonuclease, it follows that the 'stalk', though short, is more stable than the internal secondary structure elements (compare Fig. 2b). The stabilization of the stalk could be due to the formation of continuous stacking with the adjacent long hairpin, as occurs in the case of the well-studied tRNA structure. The minus strand of RQ120 RNA appears to possess essentially the same set of the secondary structure elements because it is attacked by single-strand-specific ribonucleases at the regions complementary to the loop structures of the plus strand (Fig. 2a). At the same time, the plus and minus strands differ in their tertiary structure, as evidenced by their clearly different mobilities in non-denaturing gels (Fig. 1b).

Comparison of the RQ120 RNA strands with known nucleotide sequences has shown that the minus strand of RQ120 is built up from two segments that are almost identical to the ~80-nt fragment from the last quarter of the coat protein cistron of the Q β phage RNA and to the 3'-terminal 33-nt fragment of *E. coli* tRNA^{Asp} (Fig. 2c). It follows that RQ120 RNA is a product of an *in vivo* recombination between phage and host genomes.

This conclusion is unexpected as it has long been believed that recombinations could not occur between RNA bacteriophage genomes¹⁶, but failure to detect the recombinants by genetic analysis might have been due to the high mutation rate in RNA genomes^{16,17}. It has been reported^{18,19} that the large fragment of eukaryotic tRNA^{Asp} presents at the 5' terminus of the animal Sindbis virus defective interfering particle RNAs (DI RNAs), but it was possible that the tRNA was acting simply as a primer for RNA synthesis^{18,20}. The possibility can be rejected in this case: (1) no tRNA and no primer are necessary to ensure RNA synthesis in the pure cell-free Q β replicase system^{13,21}; (2) no small replicating RNA sequenced so far other than RQ120 displays homology with a tRNA; (3) no homology exists between the tRNA^{Asp} and Q β RNA in the region adjacent to that presented in RQ120 (compare Fig. 2c). Meanwhile, the similarity in the structures of RQ120 and Sindbis DI RNAs is striking. Although it is not clear whether the participation of just the tRNA^{Asp} in both these cases is essential, the identity of the manner in which the tRNA is attached to the viral genome-derived sequence (in both cases, 5'-truncated tRNA has its 3'

* To whom correspondence should be addressed.

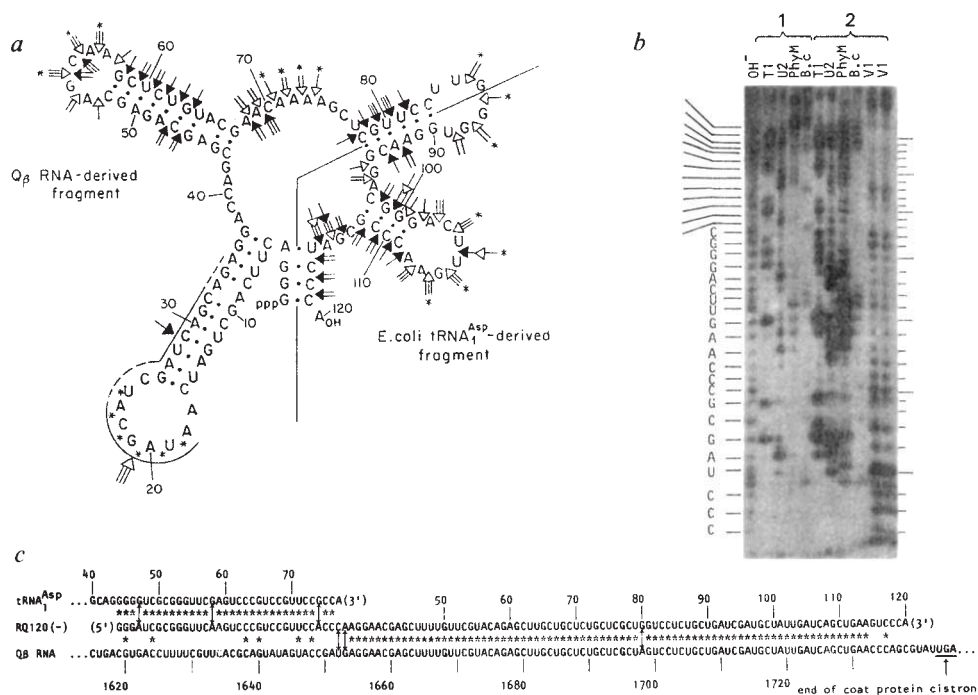


Fig. 2 Structure of RQ120 RNA. The separated labelled individual strands (Fig. 1b) were directly sequenced using chemical¹⁰ and enzymatic^{11,12} methods by >60 nucleotides beginning from the 3' end, the rest of each sequence being derived from the 3' half of the complementary strand. It has been assumed that both strands begin with pppG... because Qβ replicase always starts template reading from the penultimate C, ignoring the 3'-terminal A (ref. 13). *a*, Secondary-structure model of the RQ120 plus strand, with indication of the sites of attack by single-strand-specific nucleases T₁, U₂, PhyM, *B. cereus* (▽) and helix-specific ribonuclease V₁ from cobra venom (▼) under non-denaturing conditions. The number of lines in the arrow tails reflects the relative susceptibility to nuclease attack. Sites of nuclease digestion of the 5'-terminal segment are shown roughly due to poor band resolution in this gel region. Asterisks indicate the sites opposite to those susceptible to single strand-specific nucleases on the complementary (minus) strand. *b*, Sequencing gel pattern showing products of ribonuclease digestion of the 3'-terminal region of the RQ120 plus strand under (1) denaturing (50 °C, with or without, in the case of *B. cereus* nuclease, 7 M urea) and (2) non-denaturing conditions. *c*, Comparison of the RQ120 minus strand sequence with those of Qβ phage RNA (ref. 14, and private communication by D. Clark & M. Billetter, Zurich University) and *E. coli* tRNA^{Asp} (ref. 15). Bases in tRNA sequence are shown in unmodified form. Arrows with two heads indicate transition-type base substitutions. The search for homologies was performed using the MicroGenie sequence software (SciSoft Inc.) and the NIH GenBank (release 52).

terminus inside the recombinant molecule) may suggest the existence of a common RNA recombination mechanism in the prokaryotic and eukaryotic cells.

There are two obvious ways in which the recombinations could occur. They may arise as a result of template switching by viral replicase^{17,18}. Alternatively, there may be some kind of ligation reaction between the pre-existing RNA fragments. No enzyme like the T4 phage RNA ligase was detected in healthy or Qβ phage-infected *E. coli* cells; however, a putative enzyme activity that could compensate for the absence of both the T4 RNA ligase and polynucleotide kinase was found in uninfected *E. coli* cells by mutation analysis^{22,23}. In this regard, a linkage between the Qβ RNA- and tRNA-derived fragments of RQ120 occurs in the loop of the 'hybrid' hairpin (Fig. 2a) that would facilitate their ligation. Of course, we cannot completely exclude the possibility of a DNA-mediated pathway. This, however, seems unlikely as no DNA synthesis step exists in the RNA phage life cycle, and no specific hybridization was detected between the phage RNA and the infected host DNA²⁴.

RQ120 RNA is an autocatalytically replicating molecule, and if an internal specific enzyme recognition site exists it must be contained there. This RNA consists only of two regions of definite origin, the susceptible one being located in the Qβ RNA far away from either the two internal replicase-binding sites²⁵ or the postulated 'recognition site'¹⁸. It seems therefore unlikely that some special recognition site inside the replicating RNA is responsible for the Qβ replicase template specificity.

In many respects RQ120 RNA resembles DI genomes of animal²⁶ and plant²⁷ viruses, so generation of DI genomes seems to be a common feature of viruses of any host specificity. At

the same time, the finding that DI genomes including RQ120 RNA may contain extended fragments of viral and host genomes raises the possibility of their functioning as *in vivo* anti-sense RNAs, insofar as the complementary strands are inevitably produced at their replication.

We hope that the future discovery of the mechanism by which the replicating RNA recombinants are produced *in vivo* will lead to development of a cell-free RNA engineering technique.

We are much indebted to Professors A. S. Spirin and V. I. Agol for fruitful discussions, to Professor M. Billetter for permission to use a fragment of the unpublished Qβ RNA sequence in our paper, and to Professor A. Palmenberg for the mutant Qβamb86 phage used in growing the infected *E. coli* cells. Helpful assistance by E. V. Chetverina and A. A. Volkov is greatly acknowledged.

Received 6 January; accepted 7 April 1988.

- Kacian, D. L., Mills, D. R., Kramer, F. R. & Spiegelman, S. *Proc. natn. Acad. Sci. U.S.A.* **69**, 3038-3042 (1972).
- Sumper, M. & Luce, R. *Proc. natn. Acad. Sci. U.S.A.* **72**, 162-166 (1975).
- Bierbricher, C. K., Eigen, M. & Luce, R. *Nature* **321**, 89-91 (1986).
- Banerjee, A. K., Rensing, U. & August, J. T. *J. molec. Biol.* **45**, 181-193 (1969).
- Mills, D. R., Kramer, F. R. & Spiegelman, S. *Science* **180**, 916-927 (1973).
- Mills, D. R., Kramer, F. R., Dobkin, C., Nishihara, T. & Spiegelman, S. *Proc. natn. Acad. Sci. U.S.A.* **72**, 4252-4256 (1975).
- Schaffner, W., Ruegg, K. J. & Weissmann, C. *J. molec. Biol.* **117**, 877-907 (1977).
- Nishihara, T., Mills, D. R. & Kramer, F. R. *J. Biochem.* **93**, 669-674 (1983).
- Blumenthal, T. *Meth. Enzy.* **60**, 628-638 (1979).
- Peattie, D. *Proc. natn. Acad. Sci. U.S.A.* **76**, 1760-1764 (1979).
- Donis-Keller, H., Maxam, A. M. & Gilbert, W. *Nucleic Acids Res.* **4**, 2527-2538 (1977).
- Simoncsits, A., Brownlee, G. G., Brown, R. S., Rubin, J. R. & Guillely, H. *Nature* **269**, 833-836 (1977).
- Bierbricher, C. K. in *Evolutionary Biology* Vol. 16 (eds Hecht, M. K., Wallace, B. & Prance, J. T.) 1-52 (Plenum, New York, 1983).
- Mekler, P. *thesis*, Univ. Zurich (1981).

15. Harada, F., Yamaizumi, K. & Nishimura, S. *Biochem. biophys. Res. Commun.* **49**, 1605-1609 (1972).
16. Horiuchi, K. in *RNA Phages* (ed. Zinder, N. D.) 29-50 (Cold Spring Harbor Laboratories, New York, 1975).
17. Holland, J. *et al. Science* **215**, 1577-1585 (1982).
18. Monroe, S. S. & Schlesinger, S. *Proc. natn. Acad. Sci. U.S.A.* **80**, 3279-3283 (1983).
19. Monroe, S. S. & Schlesinger, S. *J. Virol.* **49**, 865-872 (1984).
20. Matthews, R. E. F. *A. Rev. Microbiol.* **39**, 451-474 (1985).
21. Blumenthal, T. & Carmichael, G. G. *A. Rev. Biochem.* **48**, 525-548 (1979).
22. Silber, R., Malathi, V. G. & Hurwitz, J. *Proc. natn. Acad. Sci. U.S.A.* **69**, 3009-3013 (1972).
23. Uhlenbeck, O. C. & Gumpert, R. I. in *The Enzymes* 3rd edn Vol. 15 (ed. Boyer, P. D.) 31-58 (Academic, New York, 1982).
24. Spiegelman, S. & Haruna, I. *Proc. natn. Acad. Sci. U.S.A.* **55**, 1539-1554 (1966).
25. Meyer, F., Weber, H. & Weissmann, C. *J. molec. Biol.* **153**, 631-660 (1981).
26. Holland, J. J. in *Virology* (ed. Fields, B. N.) 77-99 (Raven, New York, 1985).
27. Hillman, B. I., Carrington, J. C. & Morris, T. J. *Cell* **51**, 427-433 (1987).

Chemical probing of homopurine-homopyrimidine mirror repeats in supercoiled DNA

O. N. Voloshin*, S. M. Mirkin, V. I. Lyamichev, B. P. Belotserkovskii & M. D. Frank-Kamenetskii

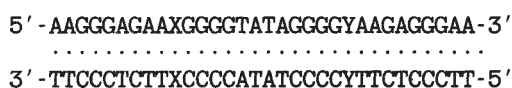
Institute of Molecular Genetics, USSR Academy of Sciences, Moscow 123182, USSR

* M. M. Shemyakin Institute of Bioorganic Chemistry, USSR Academy of Sciences, Moscow, USSR

We have recently shown that under superhelical stress and/or acid pH the homopurine-homopyrimidine tracts conforming to the mirror symmetry (H palindromes) form a novel DNA structure, the H form¹⁻⁴. According to our model, the H form includes (1) a triplex formed by half of the purine strand and by the homopyrimidine hairpin and (2) the unstructured other half of the purine strand. We used four specially designed sequences to demonstrate that, in accordance with our model, the mirror symmetry is essential for facile formation of the H form detected by two-dimensional gel electrophoresis⁴. Here we report that, under conditions favouring the H-form extrusion, guanines of the 3' half of the purine strand are protected against alkylation by dimethylsulphate, whereas adenines of the 5' half of the purine strand react with diethyl pyrocarbonate. These data indicate that the 3' half of the homopurine strand is within the triplex whereas the 5' half is unstructured, in full agreement with our model.

Dimethylsulphate (DMS) and diethyl pyrocarbonate (DEP) are widely used as probes of DNA structure because, among other reactions, they modify the N7 position of purines leading to chain scissions after treatment with piperidine. As a result the modification sites were localized at nucleotide resolution⁵. The two probes differ in specificity with respect to the N7 position. DMS reacts with guanine both in single-stranded and in duplex DNA⁵. DEP reacts mainly with adenines in single-stranded DNA but does not modify the B-form duplex, though it does modify the Z-form duplex^{6,7}. DEP was also used to detect single-stranded loops in cruciforms^{8,9}. Neither probe should react with the pyrimidine-purine-pyrimidine triplex because the N7 position of purines is completely sheltered by the Hoogsteen pairing.

According to our model of the H form²⁻⁴, guanine in the triplex is expected to be protected from the reaction with DMS. DEP is expected to modify only those adenines which belong to the single-stranded loop. We checked these predictions using four pXY32 plasmids (ref. 4) carrying the inserts:



Here X, Y are purines in the top line and complementary pyrimidines in the bottom line. We have studied all four variants: pAA32, pAG32, pGA32 and pGG32.

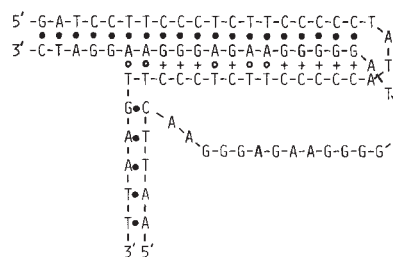


Fig. 1 The H-form structure of the pGG32 plasmid. The major element of the structure is a triplex which includes the Watson-Crick (●) duplex associated with the pyrimidine loop through Hoogsteen base pairing (○, +). One protonation site is created per Hoogsteen C-G pair (between N3 of cytosine and N7 of guanine) with the pK value near 8 (see ref. 3). The data on chemical probing presented in Fig. 2 and elsewhere indicate that of two possible isomeric forms⁴ the form shown carrying the triplex at the 3' end of the purine strand always prevails.

The H-form structure of the pGG32 plasmid is shown in Fig. 1. Our data are summarized in Fig. 2.

There is a remarkable similarity in the results for all four plasmids. The 3' half of the purine strand is protected against DMS, whereas the 5' half of the purine strand is accessible to DEP. These data were obtained for native samples of supercoiled DNA. We observed neither protection against DMS nor modification by DEP for linearized DNA samples (data not shown).

The data in Fig. 2 indicate that, in the example studied, one isomeric form prevails over the other. This isoform is presented in Fig. 1. It carries the triplex at the 3' end of the purine strand. The clear-cut protection of only half of the purine strand from DMS was first observed by Pulleyblank *et al.*¹⁰ for the 45 base-pair (bp)-long insert d(T-C)_n · d(G-A)_n (see Fig. 4 in ref. 10). But this asymmetry of the modification pattern could not be explained within the framework of their altered duplex model. Our model of the H form²⁻⁴ (see Fig. 1) explains this asymmetry in terms of different energies of the two isoforms. The data in Fig. 2, as well as that of Pulleyblank *et al.*¹⁰, indicate that the isoform with the triplex at its 3' end (as in Fig. 1) always prevails. The reason for this is unknown.

The data in Fig. 2, along with the results obtained for the same plasmids by two-dimensional (2-D) gel electrophoresis, indicate that probing with DEP is a much more sensitive technique for detecting the B-H transition than 2-D gel electrophoresis and probing with DMS. Indeed, 2-D gel electrophoresis fails to detect any transition in the case of the pGA32 plasmid (see Fig. 1 in ref. 4), whereas the pattern of modification by DEP for this plasmid at pH 4.75 is virtually as clear as in the other three cases (see Fig. 2).

There are several reasons for the greater sensitivity of the chemical (and enzymatic) probing when compared with 2-D gel electrophoresis. Chemical and enzymatic probing is applied to DNA with native superhelical density. For technical reasons, the B-H transition cannot be detected in these samples by 2-D gel electrophoresis. Moreover, to be observed by this technique, the H form should be carried by a considerable fraction of DNA molecules in the sample. The same is true for the case of DMS treatment. By contrast, probing by DEP and by S1 endonuclease give a positive result even if only a tiny portion of DNA molecules in the sample have transiently adopted the H form, because only cut molecules are detected when radiolabel is used. Also, these data are insensitive to the possible formation of other alternative structures (such as cruciforms or the Z form) in other parts of the molecule.

In the case of the pAA32 plasmid the DEP probing leads to the characteristic pattern of the H structure, even at pH 8.2. This was not observed for the three other cases studied. All experiments involving chemical modification were performed in a

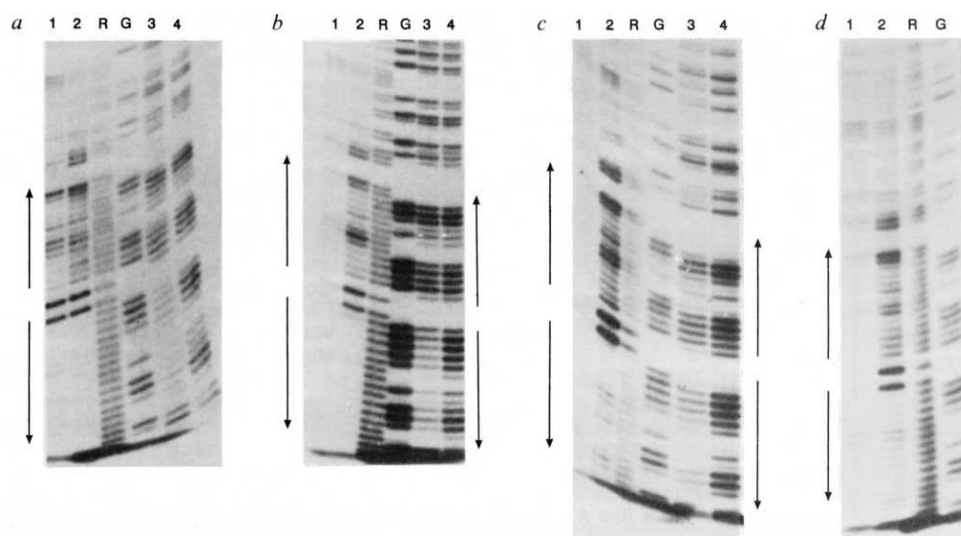


Fig. 2 Results of purine strand modification by dimethylsulphate (DMS) and diethyl pyrocarbonate (DEP). The 5' ends of the fragments are at the top. Lanes R and G are standard Maxam-Gilbert probes for purines and G. The arrows show two halves of H palindromes. The gap between the arrows indicates the 5'-TATA box. *a*, the pAA32 plasmid. DNA modification by DEP at pH 8.2 and 4.75 (lanes 1 and 2). DNA modification by DMS at pH 4.75 and 8.2 (lanes 3 and 4). Similar results are shown for the pAG32 (*b*) and pGG32 (*c*) plasmids. *d*, the pGA32 plasmid. DNA modification by DEP at pH 8.5 and 4.75 (lanes 1 and 2). The pGA32 plasmid was not probed by DMS.

Methods. We used the plasmids prepared previously⁴. Chemical modification of 3–5 µg plasmid DNA with native superhelical density was done in 200 µl of the buffer solution in the presence of 1 mM EDTA. At pH 8.2 we used 20 mM Tris HCl, 50 mM NaCl. At pH 4.7 we used 50 mM Na-acetate. We added either 2 µl of the 10% ethanol solution of DMS (v/v), incubation time 3 min at 20 °C or 2 µl of DEP (Fluka), incubation time 20 min at 25 °C. The reaction was terminated by DNA precipitation from ethanol. After reprecipitation and washing with ethanol, DNA was digested by the *Hind*III restriction endonuclease and 3'-end-labelled with the Klenow fragment of DNA polymerase I and digested by the *Pvu*II restriction endonuclease. The smaller fragment was extracted from 8% (30:1) polyacrylamide gel and treated with piperidine. The products were separated in 10% polyacrylamide gel saturated with urea. Due to poor quality of the *Hind*III enzyme, some heterogeneity of the samples was observed in lanes 3 and 4.

similar way. In the case of DEP probing the trace bands outside the inserts have similar intensities for all the plasmid studied (Fig. 2). Thus, the pAA32 plasmid was the only one to show the H form at pH 8.2. Our data on 2-D gel electrophoresis show that the superhelical density of the B-H transition for this plasmid is strongly pH-dependent at acid pH when the H form is protonated⁴. Our DEP data indicate that high enough superhelicity can induce, at least transiently, the unprotonated H structure (at pH > pK, see refs 1 and 3). The DMS data show that at pH 8.2 only a small portion of the DNA molecules carry the H form.

Evans and Efstratiadis¹¹ probed the structure of a homopurine-homopyrimidine insert with both DMS and DEP (the regular (G-A)₃₈ sequence). But, in disagreement with our data and that in ref. 10, they did not find any protection against DMS. In the case of DEP, they found the adenine residues of the insert to become hyperreactive at pH below 7, clearly indicating that an unusual conformation was formed at low pH. Also, their data clearly indicate that the 5' half of the purine strand is preferably modified by DEP in the case of non-superhelical samples at pH 4.5 (see Fig. 7 in ref. 11). This confirms the above rule. But, in disagreement with our observations, for a superhelical sample they observed modification throughout the insert. A plausible explanation for this is that in the case of a very long regular homopurine-homopyrimidine insert the second isoform is present under superhelical stress with enough probability to make the 3' half of the purine strand accessible for DEP.

Figure 2 also shows that, although no DEP modification of the homopurine strand is observed inside the triplex, at pH 4.75 the modification of the adenine residues extends up to 5 bps outside the insert adjacent to the single-stranded region of the H form. This is explained in terms of stable or transient unwinding of the AT-enriched duplex region (with the 5'-AATTC

sequence, see Fig. 1) from its open end under superhelical stress¹². The effect is not observed at pH 8.2 even for the pAA32 plasmid, which exhibits the characteristic H-form pattern of DEP modification. This suggests that the characteristic pattern at pH 8.2 reflects only a transient formation of the unprotonated H form which is nevertheless detected by the very sensitive DEP probing.

In conclusion, under acid pH and superhelical stress our pXY32 plasmids form the H structure shown in Fig. 1. This is so not only for the perfect H palindromes (X = Y) which exhibit a facile transition to the H form⁴, but also for imperfect H palindromes (X ≠ Y). We believe that the same is true for any H palindrome which is long enough and not too imperfect. This is supported by our recent observations, to be published elsewhere, of the patterns of chemical modification similar to those in Fig. 2 for a number of other H palindromes, both regular and irregular.

We thank Y. S. Lazurkin, V. G. Nikiforov and E. D. Sverdlov for discussion.

Received 22 December 1987; accepted 14 April 1988.

1. Lyamichov, V. I., Mirkin, S. M. & Frank-Kamenetskii, M. D. *J. biomol. Struct. Dyn.* **3**, 327–338 (1985).
2. Lyamichov, V. I., Mirkin, S. M. & Frank-Kamenetskii, M. D. *J. biomol. Struct. Dyn.* **3**, 667–669 (1986).
3. Lyamichov, V. I., Mirkin, S. M. & Frank-Kamenetskii, M. D. *J. biomol. Struct. Dyn.* **5**, 275–282 (1987).
4. Mirkin, S. M. *et al. Nature* **330**, 495–497 (1987).
5. Maxam, A. M. & Gilbert, W. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560–564 (1977).
6. Johnston, B. H. & Rich, A. *Cell* **42**, 713–724 (1985).
7. Herr, W. *Proc. natn. Acad. Sci. U.S.A.* **82**, 8009–8013 (1985).
8. Scholten, P. M. & Nordheim, A. *Nucleic Acids Res.* **14**, 3981–3993 (1986).
9. Furlong, J. S. & Lilley, D. M. *J. Nucleic Acids Res.* **14**, 3995–4007 (1986).
10. Pulleyblank, D. E., Haniford, D. B. & Morgan, A. R. *Cell* **42**, 271–280 (1985).
11. Evans, T. & Efstratiadis, A. *J. biol. Chem.* **261**, 14771–14780 (1986).
12. Vologodskii, A. V., Lukashin, A. V., Anshelevich, V. V. & Frank-Kamenetskii, M. D. *Nucleic Acids Res.* **6**, 967–982 (1979).

Battle of the DNA sequencers

There are now three automated DNA sequencers on the market — from Du Pont, Applied Biosystems, and EG & G Biomolecular. Below each company tells why theirs is best.

Du Pont's Genesis

Edward Chait

THE Genesis 2000 instrument from Du Pont employs the well-proven Sanger method of primer extension for DNA sequencing, but uses a unique fluorescent label to replace the traditional radioactive label for the detection of the DNA fragments. Up to 10,000 bases per day can be sequenced with one instrument, without autoradiography. The Genesis chemistry uses carefully selected fluorescent dyes attached to the dideoxynucleotides which are the terminators of the primer extension reactions in the Sanger chemistry. Thus, as each fragment is terminated by the incorporation of the dideoxynucleotide, it is fluorescently labelled and can be detected as it is separated on an electrophoresis gel. Because each one of the dideoxynucleotides (A, C, T, and G) carries its own fluorescent label — one of four dyes emitting in the yellow-green region of the spectrum — it is readily detectable from its unique spectral signature. One benefit of this method is that only truly terminated sequences are detected, eliminating "false stop" artefacts from the sequence. Another benefit is that each reaction corresponding to an unknown DNA template can be carried out in a single tube and loaded on a single lane of an electrophoresis gel, because each dideoxynucleotide will be incorporated into its proper position and can be differentiated by its dye label.

Sequencing systems which use a fluorescent dye label on the primer for the detection of the DNA fragments have the disadvantage of inflexibility because either prepared, dye-labelled, universal primers must be used, or custom-labelled primers must be labouriously synthesized. The Genesis chemistry can use any primer because the fluorescent label is on the terminator nucleotide. This allows the powerful techniques of "gene walking" to be used in large sequencing projects where the next primer can be created from the previously determined sequence. Any of the well-developed strategies for Sanger sequencing of single-stranded or double-stranded DNA may be used with the Genesis chemistry. The dye-labelled terminators are compatible with Sanger protocols using a variety of enzymes including T7 DNA polymerase.

The Genesis 2000 was designed as a total system of molecular biology, chemistry, fluorescent labels and optical detec-

tion. The fluorescent dyes were carefully chosen from a family of succinylfluoresceins which could be varied in emission wavelength without appreciably altering the molecular weight, which could cause mobility problems on the separation gel. The dyes were chosen for their high fluorescent sensitivity and their ability to be excited by the argon ion laser (488 nm) used in the instrument, to assure an optimum signal-to-noise ratio for sequencing microgram quantities of DNA.

Detecting the four emission signals from the fluorescence of each of the dye labels requires a special optical system. As separated fragments produced by the Sanger sequencing reactions proceed down the gel, for any of the twelve samples which can be loaded on the gel, they are scanned by a laser beam directed by a mirror which is the single moving part of the Genesis system. The laser excites the fluorescent emission, which is detected simultaneously by two photomultiplier tubes (PMT). In front of each tube is a filter stack: a high-wavelength pass filter for one PMT and a low-wavelength pass filter for the other. By viewing a ratio of the high-pass and low-pass signals of the spectra of the four labelled nucleotides, a unique spectral signature is obtained which allows each terminated DNA fragment to be detected, and the terminating base to be positively identified.

Once the fragments have been detected, data is recorded and processed by the Macintosh II computer which is part of the Genesis system. As a sequence run progresses, data for any one of the twelve samples can be obtained in real time. In addition, the computer controls all of the operating parameters of the instrument and monitors its functions. The advanced communication capabilities of the Macintosh II allow for the transmission of data to established databases and the transmission of data to laboratory personnel engaged in special applications.

The potential of the Genesis systems extends beyond sequencing to a host of other applications. The Genesis is a major innovation in the marriage of biology and computer science which will influence biotechnology in the future. The Genesis 2000 costs \$90,000 (US). □

Edward Chait is at the Biotechnology Systems Division, Medical Products Department, Du Pont, Wilmington, Delaware 19898, USA. Reader Service No. 100.

EG & G's Acugen

Guy Page

THE Acugen 402 is a DNA sequencer based on the detection of ^{32}P -labelled DNA fragments. The instrument uses radiolabelled DNA fragments prepared by conventional methods and inexpensive, readily available ^{32}P -deoxynucleoside triphosphates. No modifications must be made to common mini-scale template procedures to use them with the Acugen 402. The instrumentation is entirely solid-state; no calibration or adjustment is necessary for operation. In effect, the Acugen 402 automates the efficient procedures of manual sequencing, while providing powerful data collection and analysis capabilities.

The patented gel cassette system is a key element of the instrument. This gel preparation system has been engineered to simplify the process of pouring a sequencing gel, and to ensure a high degree of uniformity and reproducibility from gel to gel. The system virtually eliminates "smiling" and the related sample mobility artefacts which commonly occur in sequencing gels.

During electrophoresis, radiolabelled DNA fragments migrate past an array of narrow lead windows machined into the rear electrophoresis plate. Beta particles from the labelled fragments are recorded by extremely sensitive solid-state radioisotope detectors located directly behind the windows. Each gel lane has its corresponding detector. Information collected by the detectors is transferred to instrument memory and simultaneously to a video monitor, so that raw sequence data is available to the researcher as soon as the instrument obtains it. This allows quick evaluation of sequence quality; any DNA sequence near the primer can be read within one to two hours.

The core technology of the Acugen 402 is the solid-state ion-implanted silicon detector. The detector material is manufactured by a process that imparts an extremely low noise and low leakage current, which allows accurate detection of the low-level signals obtained from DNA sequence reaction products. The detectors are sufficiently sensitive to detect DNA molecules radiolabelled using commercially available radionucleotides and conventional techniques. Base identification is made by lane position, as with an autoradiogram, because all of the reaction products are labelled

with ^{32}P . Six reaction sets may be loaded in each run, and each run takes about six hours. Up to 3,000 bases may be sequenced per day, in two runs.

After the electrophoretic run is finished, raw data is transferred to a PC for analysis by EG & G Biomolecular's GenQuest software. The program uses a number of routines to derive the DNA sequence. These include analysis of simple peak patterns, use of the intrinsic characteristics of the data set, and use of a set of expert rules applicable to local anomalies. Sequence calculation is entirely automatic and takes a few minutes for each set of reactions. Any unresolvable ambiguities are flagged. This is an extremely important feature, because the level of variation that occurs in DNA sequence data virtually guarantees that any extensive search of a DNA sequence will reveal regions that are difficult to interpret, either manually or automatically. Because the raw data obtained from the Acugen 402 can completely be interpreted by the researcher — exactly in the manner of an autoradiogram — such ambiguities may be evaluated visually, and the sequence edited accordingly. The opportunity to incorporate the subjective interpretation of the researcher is a powerful component of fully accurate sequence analysis.

Priced at \$45,000 (US), the Acugen 402 offers fully-automated DNA sequence analysis capability at low cost. Furthermore, because the detection system is based on the use of inexpensive ^{32}P -nucleotides, the long-term operating costs are minimized. The compact design, reliable solid-state engineering, ease of use and powerful data handling capabilities make the Acugen 402 an ideal instrument for the general molecular biology or biotechnology research environment. □

Guy Page is at EG & G Biomolecular, 4 Tech Circle, Natick, Massachusetts 01760, USA. Reader Service No. 101.

Applied Biosystems

Michael Hunkapiller

APPLIED Biosystems pioneered the development of fluorescence-based, automated DNA sequencing with the introduction in May 1986 of the model 370A DNA sequencer. The key to this automation is the use of four distinct fluorescent dyes, each paired with one of the four base sequencing reactions, rather than the single radioactive label used in conventional DNA sequencing methods. The four dyes allow all of the reaction products to be separated, detected and analysed simultaneously on a single lane of a polyacrylamide gel. This coelectrophoresis eliminates most of the effects of gel distortion on the analysis of sequencing data which is often experienced in collecting results from the four or more lanes required

in conventional methods. The concept of using a set of four dyes in DNA sequencing originated in the laboratory of Leroy Hood at the California Institute of Technology. It has been developed jointly over the last several years by his laboratory and Applied Biosystems.

A key feature of the model 370A is its adaptability and versatility. The instrument is compatible with all Sanger dideoxy sequencing strategies. Inexpensive, fluorescence-labelled sequencing primers are available from Applied Biosystems. Custom primers are easily synthesized and labelled by 370A customers using a variety of methods supported by Applied Biosystems. Protocols for single- and double-stranded templates are available, along with methods for the use of avian myeloblastosis virus (AMV) reverse transcriptase, the Klenow fragment of *Escherichia coli* DNA polymerase I, modified T7 polymerase and *Thermus aquaticus* (Taq) DNA polymerase.

Results with Taq polymerase sequencing protocols developed at Applied Biosystems suggest that this enzyme may be the polymerase of choice for automated and manual sequencing. This enzyme, like modified T7 polymerase, gives much more uniform band intensities than Klenow or AMV reverse transcriptase. Unlike any of the others, the Taq enzyme works best at 70 °C, a temperature at which many of the common problems experienced in DNA sequencing are virtually eliminated. Secondary structure in the DNA template that could block the function of the polymerase is melted away, resulting in longer sequencing runs even with difficult templates. Initiation of chain extension at other than the desired site (through secondary primer binding) is also suppressed. This, along with the minimization of premature stops caused

by the secondary structure, produces sequencing data with less interfering background signals. The use of the Taq enzyme opens up the possibility of performing direct sequencing — without purification — of DNA produced by the polymerase chain reaction technique.

The model 370A is the only commercial DNA sequencer that provides individual spectral discrimination of all four dyes used in the labelling reactions. As a result, it is able to detect regions of anomalous band mobilities around GC compressions and palindromes in which bands may comigrate on the gel. Multiple, superimposed sequences from impure templates are also clearly visible.

Since the first production instrument was installed in March 1987, more than 100 instruments have been installed worldwide. During that time, the length of sequence which can be read by the model 370A has increased from 300 to greater than 500 bases, the accuracy of the automated sequencing assignments made by the system software has increased from 95 per cent to greater than 99 per cent, and the amount of template required has decreased from 6 µg to 1 µg. Current development efforts are focused on improving these performance criteria further and on increasing system throughput through shorter electrophoresis times and increases in the number of templates (currently 16) that can be analysed in a single gel run. Software to merge the sequencing data from large numbers of individual templates and to allow the use of the model 370A in restriction fragment mapping is also under development. The model 370A costs \$92,500 (US). □

Michael Hunkapiller is at Applied Biosystems, Foster City Biosystems Division, 777 Lincoln Centre Drive, Foster City, California 94404, USA. Reader Service No. 102.

ADVERTISEMENT



For research use

Introducing:

Hyaluronic Acid Test

Hyaluronic acid (hyaluronan) is present in most tissues but is especially abundant in loose connective tissues. It is released and transported via the blood to the liver where it is catabolized. Increased levels have therefore been de-

ected in patients with e.g. rheumatoid arthritis and liver cirrhosis.

Pharmacia offers you a sensitive, specific and reproducible radioassay for measurement of hyaluronic acid.

For research use only. Not for use in diagnostic procedures.

Pharmacia Diagnostics
Division of Electro-Nucleonics, Inc.
Attn: Gary Britton
368 Passaic Avenue
FAIRFIELD, NJ 07006, USA
Phone 1-800-526-3625

Pharmacia Diagnostics AB
Attn: Ulf Almqvist
S-75182 Uppsala
Sweden
Phone 46-18-16 30 00



Pharmacia

A glimpse of Achema 88

Achema 88 — the international meeting on chemical engineering and biotechnology — takes place next week in Frankfurt, West Germany. Here is a sampling of the exhibits.

WILFRIED Heil Mikroelektronik, exhibiting in hall 5.1 stand 19, calls its new instrument for the **documentation of electrophoretic gels** "the electronic equivalent of an instant photographic camera" (*Reader Service No. 103*). The purpose behind the company's electronic camera is to allow research laboratories to cut down on the use of expensive photographic materials. The camera employs an image sensor consisting of a low-noise CCD chip, which is thermoelectrically cooled to allow for longer light integration periods. From one to 500 television frames may be integrated on the sensor; the number of frames is selected by push buttons on the camera-control unit. The integrated image is stored in a frame buffer and displayed on a television monitor, and copies of the image may be printed out on high-resolution thermal paper as needed. Wilfried Heil Mikroelektronik says the resolution of the image printout is comparable to that of an instant photograph, at one-tenth the cost. The camera was specifically developed for the photography of low-light fluorescent patterns, such as those exhibited by ethidium bromide-stained DNA gels, but it may also be used for the documentation of stained protein gels or TLC plates. A complete system, including camera, ultraviolet transilluminator and photographic repro stand costs DM18,000 (West Germany).

The highlight of the J. Bibby Science Products stand will be the TemPen, a **pen-sized digital thermometer** small enough to fit into the pocket of a lab coat or jacket (*Reader Service No. 104*). The £39.50 (UK) TemPen measures just 160 × 15 × 10 mm and weighs only 50 g. It has a



Bibby's TemPen fits in a labcoat pocket.

Teflon PFA-coated probe which resists acids and alkalis and does not corrode like the stainless steel probes used in other types of digital thermometers, says Bibby. The temperature range of the TemPen is -30 to +150°C, and Bibby says it has a resolution of 0.1°C with an accuracy of ±0.5°C. J. Bibby Science Products will be in hall 6.2, stand C19-20.

A new series of **ovens and incubators** will be the centrepiece of the W.C. Heraeus stand, number FG-16-22 in hall 6.2 (*Reader Service No. 105*). Heraeus' series 6000 includes heating and drying



Heraeus' array of ovens and incubators.

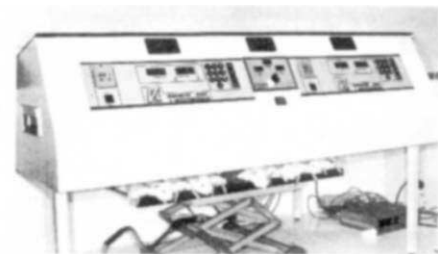
ovens, hot air sterilizers and microbiological incubators. Each incubator type comes in six different model sizes, with working space volumes ranging between 30 litres and 740 litres. All units, except biological incubators, feature natural or forced convection and are available with a manifold and other practical accessories. Electronic controllers with temperature indicators and integrated interfaces are offered with many versions. A specially-designed air circulation system which has a preheated fresh air supply improves the technical efficiency of the units and lowers the time required to heat them up. The units are also well insulated: Heraeus says the temperatures of the units' outer walls do not exceed 55°C, even when the interior temperatures reach 300°C.

Chemical Design will have its range of **molecular modelling software** on display in hall 2.0, stand D4. Chemical Design has just announced that its Chem-X molecular modelling system now offers a transparent interface to AMBER, a leading molecular dynamics simulation program (*Reader Service No. 106*). The interface allows AMBER's advanced molecular mechanics and molecular dynamics calculations to be run on macromolecular systems constructed using Chem-X. The data for

structures built in Chem-X are automatically converted to the correct format before transfer to AMBER, overcoming data input problems and errors. Chemical Design has enhanced its graphics facilities to allow dynamic simulations to be played back in real time, and soon it will be possible to run AMBER on Chemical Design's transputer-based MITIE super-workstation. Version 3.0 of AMBER, written by staff at the University of California at San Francisco, is also available from Chemical Design.

Testing and monitoring

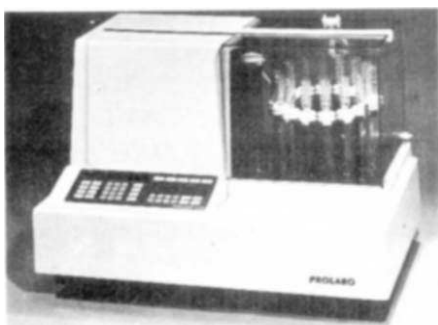
Vilber Lourmat, a French company exhibiting in stand FO11-14 in the Forum, has recently announced the introduction of its **Biotronic system for the ultraviolet irradiation of laboratory animals** (*Reader Service No. 107*). The company says the FF48,000-62,000 (France) Biotronic may be used for a range of applications, from testing the photosensitizing and photoprotective effects of cosmetic products to the study of sunburn kinetics using nude mice. The Biotronic has an irradiation aperture of 800 × 80 mm, providing sufficient space for the simultaneous treatment of up to ten albino guinea pigs, and the temperature at test level never exceeds 30°C, protecting the animals from thermal injury. An integrating radiometer control-



UV-irradiation from Vilber Lourmat.

led by a microprocessor allows precise quantitation of the radiation dose so exactly the same amount of energy can be delivered in every test. Radiation at both 365 nm (UVA wavelength) and 312 nm (UVB wavelength) is provided by three automatically selectable tube sets.

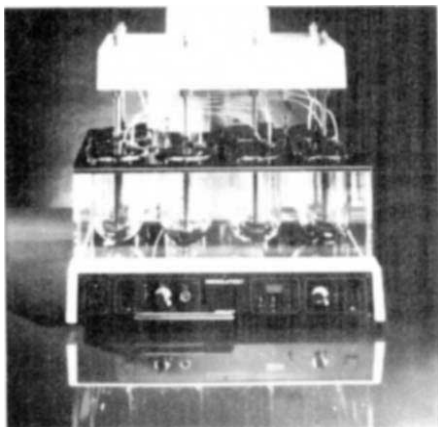
Prolabo, a subsidiary of Rhône Poulenc, will be exhibiting in stand C6-7, hall 6.2. Prolabo has recently introduced its **Microdigest A300**, a totally automatic **microwave digester** (*Reader Service No. 108*). The FF150,000 (France) Microdigest A300 can be used to carry out the mineralization of organic compounds by the wet method, and for the preparation



Prolabo's Microdigest A300 microwave digester.

of mineral solutions to determine their content. Prolabo says the Microdigest A300 can mineralize thermoplastic materials such as PCV, polyethylene and polyamide, and biological products such as blood and urine samples, in 15 minutes. Food products, including meat and milk, can be mineralized in 15 to 30 minutes. The Microdigest A300 is equipped with a sample carousel which can accommodate 16 30-ml samples. The power, time, and injection of four possible reagents can be programmed in either English or French.

May & Baker, exhibiting in stand J8A, hall 6.2, sells the dissolutes **dissolution tester** (*Reader Service No. 109*). The Dissolutes dissolution tester is a seven-stand



A dissolution tester for the pharmaceutical industry.

apparatus for the testing of tablets or capsules in pharmacotechnics laboratories. The Dissolutes can process six samples concurrently, with the seventh vessel being used as a cool medium reservoir to serve as a reference or as a reservoir for HPLC applications. Each of the dissolution vessel stands is fitted with a specially-designed cover to prevent evaporation loss, and the bath has duplicate thermostats with double safety switches. The reactor vessels are made of thick borosilicate glass, and the unit is built in one block to eliminate vibration.

The model MA50 **moisture analyser** is the latest instrument off the production line of Sartorius, exhibiting in stand KL21- 24, hall 6.3 (*Reader Service No. 110*). The MA50 is designed for both

moisture analysis in the laboratory and for quality control in the industrial laboratory. Parameters and settings for running up to 10 programs can be entered and stored for a variety of samples. Unauthorized access or accidental changes to the settings are protected by a password requirement. The built-in printer provides a hard copy of the analytical sequence, and printouts can also be obtained for the

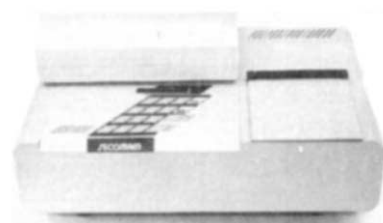


Moisture analysis from Sartorius.

last 20 analyses. Sartorius says the MA50 works with the precision of an analytical balance, and can provide moisture determinations for samples weighing as little as 200 mg. Minimum moisture levels of 0.1 per cent or less can be determined with the instrument, with 0.1 mg accuracy, says the company.

The analytical tradition

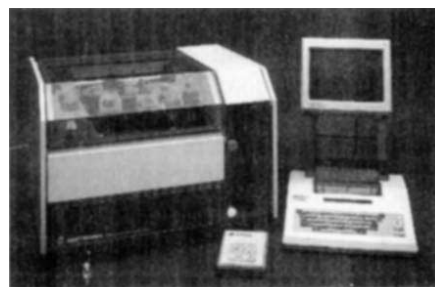
Secomam has a new **spectrophotometer** for applications in pharmaceutical, food development, chemical, cosmetic and environmental laboratories, and teaching institutions (*Reader Service No. 111*). The model S.500.1 is available for \$3,200 (US), and Secomam says it is the only spectrophotometer in its price range to include both endpoint and kinetic modes of measurement with automatic multi-standard calibration. The S.500.1 is microprocessor-driven, and can keep up to 24 user-defined protocols and labels in permanent memory. The software automatically manages all functions including wavelength selection, automatic zeroing,



Secomam's multi-purpose spectrophotometer.

automatic reagent blanks and standard calibration. The two measurement modes are computed automatically, using factor or multistandard calibration with one to five standards. Two reading modes are displayed in absorbance or transmittance. Secomam will be exhibiting in stand DE4, hall 6.3.

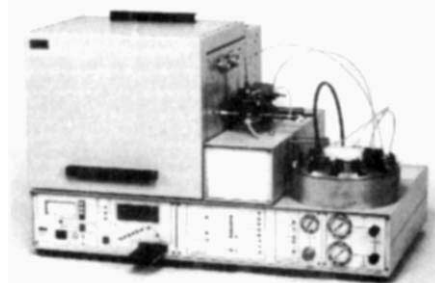
Biotronik has over 800 units of its new model LC3000 **amino acid analyser** in the field (*Reader Service No. 112*). The computer-controlled model LC3000 is a system for the post-column detection of amino acids after separation on an ion-exchange column, and is geared toward the routine user. The instrument can be automatically or manually controlled, and has built-in self-diagnostic systems. Menus and easy-to-read graphics allow user-friendly programming. Up to 20 separation programs can be stored in computer memory, and 30 canned programs are available from Biotronik's program library. The LC3000 has an automatic sampling system in a cooled cabinet which can hold 200 micro vials or 100 standard vials. Pre-packed columns



Biotronik's computerized aa analyser.

for hydrolysate or physiological fluids and ready-made buffers are available. Biotronik will be in stand DE2-3, hall 6.2.

The French company Gira, exhibiting in stand DE2, hall 6.3, has a **preparative**



Preparative GC from Gira.

gas chromatograph capable of fully automatic operation — the CAP.12 (*Reader Service No. 113*). Developed in conjunction with Rhône Poulenc, the CAP.12 has an adjustable injection system which Gira says has excellent reproducibility even with very small sample volumes. The 12 glass traps of the CAP.12's rotating plate collector feature a patented design. The FF200,000-260,000 (France) CAP.12 can be used preparatively with filled columns and a thermal conductivity detector, or in

a micropreparative mode with capillary columns and either a microthermal conductivity detector or a flame ionization detector with an output divider. Analytical separations are also possible.

In the Leitz Wetzlar stand number KL14-15, hall 6.3, Heidelberg Instruments will be showing off its **3-dimensional laser scanning microscope** (Reader Service No. 114). The Heidelberg Instruments microscope allows the 3-dimensional measurement and quantitative analysis of structures in confocal, non-confocal or multi-colour fluorescent modes. It can be used to show regenerative processes, fix chromosomes in cell nuclei, determine cell volume, examine living preparations, and measure surfaces in materials technology. Heidelberg Instruments says the microscope can resolve and measure structures down to 0.2 μm , and has the capability to present vertical optical image sections. The system has both a HeCd laser and an Ar laser, which can be used separately or in combination. Up to eight images per second can be imaged in various matrices, to a maximum of 1024×1024 pixels, says the company. Relevant software is included, and can be expanded by user programming.

Dispensing

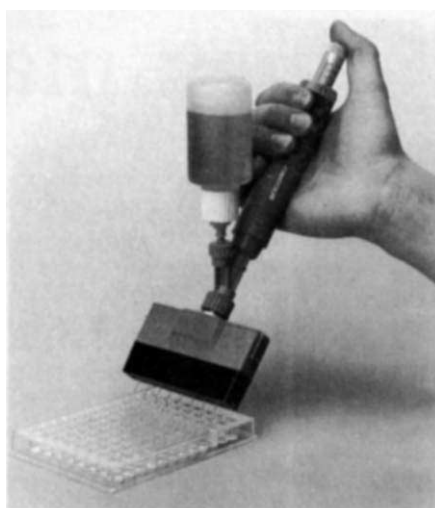
Jouan has recently introduced its model PM600 "intelligent" peristaltic pump (Reader Service No. 115). Jouan says the \$1,600 (US) pump is particularly well adapted for repeated dose delivery of viscous liquid agar media in conjunction with the Jouan model SH media prepara-



Filling Petri dishes is just one of the uses of Jouan's peristaltic pump.

tor. The pump has a microprocessor that allows it to be programmed for volume and cycle, semi-automatic or security modes. Ten permanent memories are available for the storage of programmed parameters. Pump speed, pause time and pump direction are also programmable, and several programs can be linked together for repetitive dispensing. The 16-digit LCD display shows the status of the pump at all times. Jouan will be exhibiting in B6-8, hall 6.3.

An **eight-channel microdispenser** is new from Socorex, exhibiting in stand J4, hall

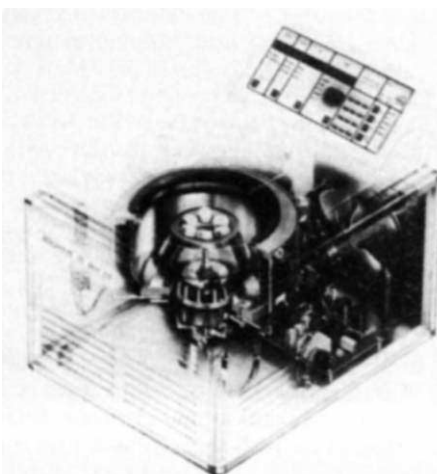


Eight channels in one pipetting is possible with the Socorex multi-channel dispenser.

6.2 (Reader Service No. 116). The model 868 multichannel dispenser can deliver liquid samples of between 10 μl and 250 μl into standard microplates with 9 mm spacing. The basic unit of the dispenser is an adjustable micropipette fitted with a three-way valve for automatic refilling and an eight-channel flow splitter. Socorex says both the deviation between wells and the C.V. are better than 2 per cent. Reagent bottles can be attached directly to the unit, or a flexible feed tube can connect the dispenser to a remote sample supply. The dispenser has an inert fluid path made of ceramic, Teflon, platinum-iridium and stainless steel, and is autoclavable to 121 $^{\circ}\text{C}$. Each dispenser comes with a calibration print-out to certify its accuracy.

Spinning down

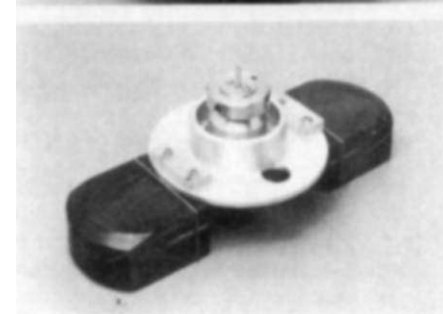
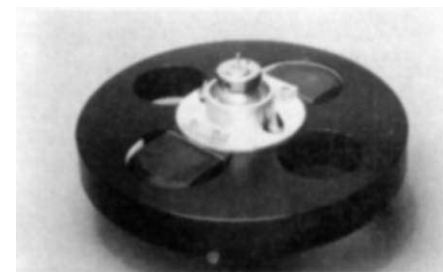
Sigma Laborzentrifugen GmbH will be exhibiting its latest **refrigerated centrifuge** in stand H16, hall 6.2 (Reader Service No. 117). The Sigma model 2K15 has a maximum speed of 15,300 r.p.m. to yield up to $20,000 \times g$, and a temperature range of between -20°C and $+40^{\circ}\text{C}$. The centri-



The model 2K15 refrigerated centrifuge from Sigma is good from the inside out.

fuge holds 6×30 ml, 10×10 ml, 12×2.0 ml and 24×2.0 ml rotors, and a magnetic rotor identification feature prevents operating errors. The model 2K15 is controlled by a microprocessor that can calculate all centrifugation parameters, and operation is simplified by separate "start" and "stop" buttons. It has a double-cover lock for security, and Sigma says its maintenance-free drive motor is extremely quiet because it lacks collector and carbon brushes.

Beckman will be bringing its newest **elutriator rotor** to Frankfurt, and will be displaying it in stand DE9-12, hall 6.2 (Reader Service No. 118). The 10,200 (US) model JE-5.0 is designed for the counterflow centrifugation and recovery of specific populations of living cells, using the Beckman J6 series of floor-model centrifuges. Beckman says its elutriator rotor can be used in synchronous cell growth studies, the separation of liver cells and of



Beckman's elutriator rotor, with a close-up of the elutriation chamber.

human lymphocytes and monocytes, and other cell and particle studies. Elutriation separates cells by a gentle, washing action in an isotonic medium. Viable cell recovery rates are close to 100 per cent for many types of cells including bone marrow, brain, lung, blood and testis, says the company. The autoclavable rotor has two chambers for the elutriation of a 5-ml sample volume for harvesting cell populations of 10^7 to 10^9 cells, and a 40-ml volume for obtaining cells numbering from 10^7 to 10^{10} . Beckman says the rotor can separate whole, living cells ranging in size from $2\mu\text{m}$ to $5\mu\text{m}$. □

These notes are compiled by Carol Ezzell from information provided by the manufacturers. To obtain further details about these products, use the reader service card bound inside the journal. Prices quoted are sometimes nominal, and apply only within the country indicated.

New skills for new materials?

Richard Pearson

Industrial application of recent advances in ceramics technology will require adaptable staff with first degrees in a variety of scientific subjects, rather than highly specialized personnel with higher degrees.

THE new technologies continue to attract the attention of policy makers and educationalists as they seek to ensure that sufficient skilled manpower is available to support their development. In the case of information technology, the number of skilled professionals is large (currently 230,000 people in the United Kingdom) and has been expanding at more than 5% per annum for over a decade. Employers look to higher education to supply recruits, and the government has recognized this by setting up two initiatives aimed at boosting the supply of new graduates¹. In the case of biotechnology the number of skilled personnel is far smaller, totalling less than 5,000 in the United Kingdom, but demand is growing by at least 10% per annum². The biotechnology labour market, which has only really developed in the 1980s, is also distinguished from that for information technology by the high educational attainment of its workers, most of whom have PhDs. In contrast, information technology is dominated by holders of bachelor's degrees, many of whom did not even study an information technology-related subject. In biotechnology the commercial payback is not expected until the late 1990s, another contrast with information technology where the returns are much more immediate.

What then of the third strand of the new technologies — new and advanced materials? Like biotechnology, which has been around for several thousand years in the guise of brewing and baking, materials such as ceramics and metals can be traced back just as far, if not further, in time. It has been the last 25 years, however, which has seen the emergence of a new range of structural materials such as advanced ceramics, polymer composites and strong engineering plastics, with sectors such as aerospace and vehicles spearheading their application. Ceramic products now range from the traditional cup or decorative vase, to new applications as diverse as wear-resistant bearings, light-deflecting video displays, artificial bones and cladding for nuclear reactors. Three broad categories of companies have been identified as being involved in the advanced ceramics sector; first, materials companies that are diversifying; second, traditional ceramics and glass companies upgrading their activities; and third, users of the new materials. There were over 50 companies of each type active in Japan in 1985, while a similar pattern was evident in the United

States with a strong user sector. In Sweden and Germany attention has focused on heat engine applications with the close involvement of the automobile sector, while interest in France built up with the nuclear energy programme in the 1960s and 1970s and has subsequently broadened.

In the United Kingdom a recent survey showed that existing firms were sceptical about a major market for advanced ceramics emerging in the medium term; thus, with few users involved in research and development, most companies fall into the first two categories. Future developments and applications were seen to depend critically on technical advances which would improve reliability and reproducibility in production, and reduce costs. These developments are not anticipated for 5–20 years and will require considerable financial and human resources in research and development. The companies' priorities are to aim for market niches or for the replacement of existing materials, rather than for new products or processes. A major concern was that there was no major UK mass-engine manufacturer giving the support and commitment that advances ceramics industries overseas are receiving³.

What of the skills involved? In research and development the key need was for high-quality graduates in basic disciplines such as chemistry, physics and chemical engineering who could then be taught about ceramics. Materials scientists and metallurgists were also employed, although in part this was due to a paucity of ceramics graduates in the labour market. A key need in several companies was for people to be able to work with, and adapt to, several different materials rather than to specialize in one. Materials scientists had often been found to be lacking in depth, being seen as "Jacks of all trades". Criticisms were also levelled at physicists for being too focussed and unable to deal with processes and end products, while chemists were often seen as good at making ceramics, but less interested in the products. What was seen as most important was people with a mixture of skills who would be highly adaptable and prepared to learn new techniques. Into the production stage, the main need was for mechanical and chemical engineering skills with the ceramics expertise being added later. A final need for technical skills was in the area of sales and marketing, where customers needed

technical data to convince them of the properties and advantages of the new materials.

In total it was estimated that the number of research workers in ceramics in the United Kingdom was less than 1,000, very small even in relation to biotechnology, and minuscule with respect to information technology. Comparable figures for Japan were 6,000 graduates plus another 6,000 technicians and other support workers. In the United States the total in ceramics research was about 2,000 engineers and 4,000 technicians.

Only one UK university in the has a department of ceramics and student numbers there are small, about 20 undergraduates per year, while the MSc course is likely to close as it attracts too few students. Other degree courses, however, have ceramics options and there are PhD students working on ceramics. In Japan there are only three ceramics courses in national universities and on-the-job training is emphasized instead. In contrast ten US universities have undergraduate ceramics degree courses and twelve provide graduate courses.

The general picture for this new technology is thus one of learning by experience, and for once it is a new area not beset by skill shortages nor apparently needing large numbers of new programmes in higher education. As such, the reliance on basic scientific and engineering disciplines enables rapid responses by higher education to changes in demand. The industry sees itself to be on a gradual growth path and there is not expected to be a sudden, dramatic increase in the demand for ceramics skills in the United Kingdom, particularly given the current dominance of Japanese and US producers in world markets. The skills base is small, however, and if potential users become more aware of the value of new materials the low availability of experienced staff could constrain their investment in research and development. Nevertheless, the new materials seem to be the least likely of the new technologies to be demanding a massive investment in higher education this century. □

Richard Pearson is at the Institute of Manpower Studies, Mantell Buildings, University of Sussex, Brighton BN1 9RF, UK.

1. Pearson, R. *Nature* 333, 100 (1988).

2. Pearson, R. *Nature* 328, 96 (1987).

3. Brady, T. *Advanced Ceramics* (Manpower Services Commission, 1988).